The role of deleterious passengers in cancer

The role of deleterious passengers in cancer

A DISSERTATION PRESENTED
BY
CHRISTOPHER D. MCFARLAND
TO
THE COMMITTEE ON HIGHER DEGREES IN BIOPHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOPHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
AUGUST 2014

Thesis advisor: Professor Leonid A. Mirny Christopher D. McFarland

The role of deleterious passengers in cancer

Abstract

The development of cancer from a population of precancerous cells is a rapid evolutionary process. During progression, cells evolve several new traits for survive and proliferation via a few key 'driver' mutations. However, these few driver alterations reside in a cancer genome alongside tens of thousands of additional 'passenger' mutations. Passengers are widely believed to have no role in cancer, yet many passengers fall within functional genomic elements that may have potentially deleterious effects on the cancer cells. Here we investigate the potential of moderately deleterious passengers to accumulate and alter neoplastic progression.

Evolutionary simulations suggest that moderately-deleterious passengers accumulate during progression and largely evade natural selection. Accumulation is possible because of cancer's unique evolutionary constraints: an initially small population size, an elevated mutation rate, and a need to acquire several driver mutations within a short evolutionary timeframe. Cancer dynamics can be theoretically understood as a tug-of-war between rare, strongly-beneficial drives and frequent mildly-deleterious passengers. In this formalism, passengers present a barrier to cancer progression describable by a critical population size, below which most lesions fail to progress, and a critical mutation rate, above which cancers collapse. In essence, cancer progression can be subverted by its own unique evolutionary constraints.

The collective burden of passengers explain several oncological phenomena that are difficult to explain otherwise. Genomics data confirms that many passengers are likely damaging and have largely evaded negative selection, while age-incidence curves and the distribution of mutation totals suggests that drivers and passengers exhibit competing effects. These data also provide estimates of the strength of drivers and passengers.

Finally, we use our model to explore cancer treatments. We identify two broad regimes of adaptive evolutionary dynamics and use these regimes to understand outcomes from various treatment strategies. Our theory explains previously paradoxical treatment outcomes and suggest that passengers could serve as a biomarker of response to mutagenic therapies. Deleterious passengers are targetable by either (i) increasing the mutation rate or (ii) exacerbating their deleterious effects. Our results suggest a unique framework for understanding cancer progression as a balance between driver and passenger mutations.

Contents

The following authors contributed to all Chapters:
Christopher D. McFarland
Leonid A. Mirny
Kirill S. Korolev

The following authors contributed to Chapter 4:
Gregory V. Kryukov
Shamil R. Sunyaev

Listing of figures

This dissertation is dedicated to the enjoyment of science.

Acknowledgments

When I first met my advisor, Leonid Mirny, he was number seven on the list of faculty that I wanted to meet during my prospective visit to Harvard. During my visit, I was struck by how knowledgeable he was of disparate fields. After matriculating, I approached him to discuss a rotation involving his work on transcription factor biophysics, but he told me that he had a more interesting idea and that cancer genomics/modeling was about to grow as a field. Ever since then, Leonid and I have been pursuing more interesting ideas. Leonid's enthusiasm, creativity, and ability to integrate various fields and ideas has always impressed me most. The ingenious of this project—connecting cancer's addiction to chaperone proteins to its evolutionary history—was conceived by Leonid before I walked into his office. Yet his support didn't stop there. It eventually took us five years to publish our first work, and while I naturally lost enthusiasm at times, Leonid never did. I cannot recall a single personal meeting, lab meeting, or lecture where Leonid's enthusiasm or attention was absent. He has perpetually encouraged me to try out new ideas and my own pet-projects, travel to more conferences, attend more classes and lectures, and to 'optimize fun'. I could not have had a more supportive advisor or better role model.

I joined the lab during a transitionary period. Its elder members, Anahita Tafvizi and Jason Leith, were largely focused on experimental work outside the lab and on graduating. They were wonderful counselors and I'll particularly enjoy Jason's enlightening perspectives on science and life that I digested after midnight in the lab.

The Mirny Lab's new guard consisted of Maxim Imakaev, Geoff Fudenberg, and myself. We formed a gang that I'll sorely miss. Early on, Max taught me nearly everything about computer hardware and systems software. We (mostly Max) assembled a beowulf cluster and lab network. Max helped me develop the computer skills I needed throughout my thesis. Geoff and I first connected over weekly runs. He would casually discussed the latest science and personal affairs, as I gasped for air. He has continually referred me to noteworthy studies that I should have already read and that Geoff should not have known about.

Anton Goloborodko, joined 'The Leonids'—our gang's epithet among outsiders, almost three years later and quickly assimilated into the group. He has been a great

friend, especially when I want to get outside, away from the glowing pixels, and is a tremendously useful resource to bounce ideas off. I think of each of my labmates as having a particular asset that I was most impressed by: Max's ability to intuit any scientific problem in under 5 minutes; Geoff's lucid, nearly poetic, writing; and Anton's rigorous, analytical approach. They were all tremendously helpful at every step in the scientific process: the review of outside literature; methodological design; troubleshooting; interpretation of results; writing; and (most critically) the celebration of successes, failures, and boredom. I would also like to mention Boryana Doyle, a precocious student who managed to infiltrate our clique and became another good friend. Most importantly, these people formed the center of my social life for six years and will make it hard for me to move on.

There are many faculty outside of The Leonids that were instrumental to this research. Most notably, Gregory Kryukov and Shamil Sunyaev provided the necessary genetic knowledge and feedback, while Martha Bulyk was a great rotation advisor and mentor, and Jacob Scott provided me a clinical perspective and excellent humor. Nevertheless, Kirill Korolev was absolutely instrumental to this project. All of this project's analytical analysis was either done directly Kirill or closely supervised by him. He became a tremendously useful resource, as I tried to learn population genetics and evolutionary theory with little formal training. Lastly, he became a great personal contact that I hope to retain throughout my scientific career.

Your first thought when reading this acknowledgment may have been: 'Prospective students to Harvard Biophysics have the option to meet with at least 7 faculty during their visit!?' In fact, I was given the option to request meetings with up to ten Harvard faculty. This was all arranged by Harvard Biophysics' exemplary administrator: Michele Jakoulov. The phrase 'Michele and Jim [Hogle, the Department Chair,] are the best thing about Harvard Biophysics,' is an adage used so often in my program that it has become a clichè. So I would like to nominate a new truism: Harvard Biophysics is a family, rather than a program, because of Jim and Michele. Like all family, their incredible effort put into the community is too frequently taken for granted. I most appreciated their friendship towards myself and everyone else in the program. I've hung out with Jim at 2 AM after a biophysics event officially ended (or at a house party) recalling memories of yore as Brad Nelms jamed on his fiddle. I've received personal palm readings from Michele, while packed into the M-2 bus like sardines. Not only do they want us to find our own ways and succeed, they have a sincere interest in our lives and well-being. They treat us as family.

As I currently write this, my longterm girlfriend Christine Miller, is sitting next to me doing her own research. We met about halfway through my PhD and she has helped me immensely since. This includes editing thesis chapters as I write them; accommodating the endless disruption of plans that my graduate experience has created; immersing herself in my academic culture and lifestyle; and most importantly, always being supportive. She is a kindred-spirit and a great companion. I particularly

admire her patience and optimism and have relied upon them to complete my own PhD; I'm certain that she will excel at her PhD.

Not only are some of the scientists I have met family, my family are scientists. My family is a major reason why I enjoy science and have chosen my current life. My dad would describe the Krebs cycle to me on camping trips when I was eight years old. Today, we continue to exchange emails about early human adaptations and gravity waves. My mom, although also a scientist, has always been more focuses on my moral and social development. She has always lectured me on the importance of family, manners, working hard, speaking up for myself, handling finances, and social etiquette. While its impossible to adequately convey everything parents do for a child in writing, they have, in essence, made me into the person I am today.

Nothing in Biology Makes Sense Except in the Light of Evolution

Theodosius Dobzhansky

1

Introduction

Cancer is a disease of clonal evolution in the body[84]. Somatic cells divide and die hundreds or thousands of times during a human lifetime. They exist within a niche of the body, often termed a *microenvironment* in cancer literature, that requires these cells to respond to stimuli. Finally, somatic cells can acquire their own somatic mutations that often confer new phenotypic traits to the cells, which may then be selected for or against by the cell's microenvironment. For these reasons, somatic tissues can

evolve. They change morphology and adapt to their environment and, as such, can only be understood 'in the Light of Evolution'[84,37,56,8,58,76,7,104,88].

The early, small populations of somatic cells that may eventually evolve into cancer, or *lesions*, were first identified by their two most salient phenotypes: an abnormally fast rate of cell division, and an abnormally large degree of variability in cell morphology[122]. This first phenotype (enhanced cell division) is an irony of evolution: enhanced cell division is highly advantageous to individual somatic cells in a microenvironment, yet highly disadvantageous (often lethal) to the organism as a whole. Thus, mutations that increase cell's rate of proliferation are selected for in somatic tissue, yet selected against in the organism's germ-line[29]. However, the second most salient feature of lesions, increased phenotypic variability, is equally important for their evolution into cancer. Indeed, without a continual source of new phenotypic variation, precancerous cells will never sample the other carcinogeneic traits necessary for progression.

Precancerous evolution, or cancer progression, is an example of a 'rapidly adapting population': cancers develop as many as ten new traits[54], exhibit a high mutation rate[54,7,72], and rapidly change in population size[30]. Developing so many new traits in an evolutionarily-short time period of approximately 10,000 cell divisions is only possible because new traits arrive at an accelerated pace, which is only possible because precancerous cells's exhibit a high mutation rate[7].

While many populations evolve new traits via a gradual accumulation of changes, some, like cancer, adapt very rapidly. Other examples include viral adaptation during infection[60]; the emergence of antibiotic resistance[128]; and artificial selection in biotechnology[124]. Rapid adaptation in all these circumstances is characterized by three key features: (i) the availability of strongly advantageous traits through rare

mutations, (ii) an elevated mutation rate[91,60], and (iii) a dynamic population size[18]. Because traditional theories of gradual adaptation are not applicable under these conditions, new approaches are needed.

Recent advances in sequencing and genotyping of cancer tissues at a genome level have found that individual cancers contain tens of thousands of somatic alterations[101,102,11,111], which can be explained by cancer's enhanced mutation rate and enhanced rate of cell division. These alterations arise in many forms, such as single-nucleotide substitutions, insertions, deletions, rearrangements, Loss Of Heterozygosity (LOH) events, copy-number alterations, and whole-chromosome duplications/deletions[101]; epigenetic alterations[107]; and inheritable changes in cell state. Progression is driven by only a handful (2-15) of cancer's myriad of mutations[71] and chromosomal abnormalities[127]. From an evolutionary perspective, drivers confer advantageous phenotypes to neoplastic cells (i.e., phenotypes that allow cells in the population to proliferate further). This property is inferred by their effect on cancer-related pathways; frequent occurrence at the same genes, loci, or pathways in different patients[11,111,85]; and by the structure of cancer incidence rates[52]. Drivers arise in cancer-related genes (oncogenes and tumor suppressors) and are beneficial to cancer cells because their phenotypes either increase the cell proliferation rate, eliminate brakes on proliferation, facilitate uncontrolled proliferation or other hallmarks of cancer[54]. Because the arrival of driver mutations is so critical to cancer progression, their discovery has been the primary goal of genome-wide cancer sequencing[39].

In contrast, the overwhelming majority of mutation events in cancer are believed to have non-significant phenotypes and are called passenger alterations or simply *passengers* (table 1.1 on the following page). Little attention has been paid to passengers[111]. These alterations are assumed to be phenotypically neutral in cancer cells be-

Table 1.1: Average number of driver and passenger mutations by tumor type. The total number of identified Somatic Nonsynonymous Mutations (SNMs) and Somatic Copy Number Alterations (SCNAs) for various tumor-normal paired sequences from various tissues of origin: 100 breast [115], 183 lung [28], 159 Colon without Microsatellite INstability (MIN$^-$), 64 Colon with Micro-Satellite Instability (MIN$^+$) [21], and 121 melanomas [10].

Cancer	NSM Drivers	NSM Passengers	SCNA Drivers	SCNA Passengers
breast	1.7	70.8	1.0	34.6
lung	2.3	348.6	8.4	89.5
colon, MIN$^-$	8.8	114.0	14.1	583.5
colon, MIN$^+$	28.8	489.0	12.7	235.1
melanoma	7.0	379.6	12.6	324.7
all	9.1	272.8	8.8	258.9
Max	28.8	489.0	14.1	583.5
Min	1.7	70.8	1.0	34.6

cause they are non-recurrent and are dispersed across a cancer genome[39,22]; however, their phenotype has never been systematically tested. If passengers arise as random alterations, then many could be deleterious to cancer cells[90,27,40], potentially via proteotoxic stress[93,41,108], direct loss of function[14,79], or immune provocation[3,120] among other possible mechanisms.

Although highly deleterious passengers are expected to be weeded out by negative selection[7], moderately deleterious passengers can evade negative selection and accumulate by mutation-selection balance, ratcheting, or similar mechanisms studied in population genetics[44]. Protein-coding passengers may individually exert small effects, but because cancer genomes contain hundreds to thousands of passengers, they may collectively be significant enough to alter the course of cancer progression.

While the role of deleterious mutations in cancer is largely unknown, their effects on natural populations has been extensively studied in genetics[77,57,48]. The accumulation of deleterious mutations can cause population extinction by a process known as

Muller's ratchet or mutational meltdown[77,92]. How this applies to rapidly adapting populations with a varying size, with highly advantageous mutations, and specifically to cancer, remains unknown.

Rapidly adapting populations face a double bind: they must quickly acquire these, often exceedingly rare, adaptive mutations, yet also avoid mutational meltdown. As a result, adaptive processes frequently fail. Indeed, less than 0.1% of species on earth have adapted fast enough to avoid extinction[82] and, similarly, only ∼0.1% of precancerous lesions ever advance to cancer[98]. Cancer therapies may be able to exploit this pitfall of evolution (extinction) and cancer's distinctive method of rapid adaptation to avoid extinction[67].

Here we investigate the possible role of deleterious passenger alterations in cancer progression and examine their potential as an unexploited therapeutic target. First, we develop an evolutionary model of rapid adaptation, where (i) population size can change freely, (ii) individual's experience an elevated mutation rate that, (iii) can generate both advantageous driver mutations and (iv)deleterious passenger mutations. We find that moderately deleterious passengers can arise alongside drivers in this model, and evade purifying selection. The accumulation of passengers alters the dynamics of cancer progression and may explain several clinical phenomena, such as slow progression, long periods of dormancy, the prevalence of small subclinical cancers, spontaneous regression, and heterogeneity in growth rates. These phenomenon cannot be easily explained without considering deleterious passengers. Unlike the current driver-centric paradigm of cancer progression, our analyses demonstrate that progression depends on drivers overcoming passengers.

Mutational meltdown is possible and common in evolving precancerous populations. In fact, we observed a tug-of-war between beneficial drivers and deleterious pas-

sengers that creates two major regimes of population dynamics: an adaptive regime, where the probability of adaptation (cancer) is high; and a non-adaptive regime, where adaptation (cancer) is exceedingly rare. These regimes are separated by an effective barrier, which makes cancer progression an unlikely event.

We then tested the model's predictions by analyzing somatic mutations sequenced in cancers. This analysis shows that, in agreement with the model, individual passengers are likely to be damaging to cells and have largely evaded negative selection. Drivers, in contrast, exhibit signatures of highly non-neutral phenotypes. Our model is consistent with cancer genomic patterns and age-incidence data, offers a new interpretation of cancer treatment strategies, and explains a previously paradoxical relationship between cancer mutation rates and clinical outcomes. Most importantly, it suggests that deleterious passengers offer a new, unexploited avenue of cancer therapy.

Thus, we used our model to explore two possible therapeutic approaches that target passengers' collective phenotype and find that increasing either the mutation rate or the deleterious effect of passengers leads to cancer meltdown. The latter therapy may be possible by targeting pathways that buffer the effects of mutations, e.g., unfolded protein response (UPR) pathways.

Finally, we present and discuss clinical and biological evidence that supports an important role of passenger alterations in cancer. Our analysis further explains how asexual populations such as cancer rapidly evolve new traits while occasionally avoiding mutational meltdown.

2

An evolutionary model of cancer progression with deleterious passengers

In this chapter, I present a very simple model of cancer progression incorporating deleterious passengers. While the model is simple, the chapter is rather slow and methodical because I want to entertain all reasonable models of cancer progression before settling on the one that constitutes my primary focus. This entails review-

ing some of the more technical properties of cancer and it corresponding models. It also requires discussing circumstances where models are equivalent to one another; and then the circumstances that are most relevant to cancer progression, which varies from tumor to tumor. So although this chapter may seem dry, I actually believe it encapsulates the most challenging aspect of science: literature review at a technical level, analytical deconstruction of theories, and developing a rigorous intuition for the strengths and limitations of existing models and paradigms.

Existing evolutionary models of cancer progression have several limitations. Many models have considered a population of a constant or externally controlled size[113,8], which does not depend on the absolute fitness of cells. Other models study exponentially growing cancer populations[52,8,56], whereas logistic-like behavior has been observed in cancer[70]. Most importantly, the vast majority of cancer models neglect the effects of passenger alterations. In the last section of this chapter, I discuss two publications that do briefly consider deleterious passengers in cancer.

2.1 Cancer as a birth and death process

We model cancer at the cellular level. Cells can divide, potentially acquiring mutations, and die. Generally, the birth and death rates of a cell depend on the effect of accumulated drivers and passengers, and the environment. Assuming that all drivers/passengers possess equal fitness advantage/disadvantage, the birth and death rates $B(n_d, n_p, N)$ and $D(n_d, n_p, N)$ of each cell depend on the number of drivers n_d, the number of passengers n_p, and the total hyperplasia or population size N. In chapter 3 on page 22, we model drivers & passengers of variable effect size. Hence, we can model these competing birth and death rates using a first-order Gillespie Algorithm, specifically the

Next Reaction[43]. Each cell within the population is represented as a separate 'chemical species' or reactant in the Gillespie algorithm. Cells are then defined by their state $\{n_d, n_p\}$ in this simplest algorithm, and by a generalized genome vector in more sophisticated implementations.

Population size changes with the birth and death of individual cells (figure 2.1 on the following page). Driver mutations increase population size by either increasing the birth rate (e.g., an activating mutation in $KRAS$) or by decreasing the death rate (e.g., a $TP53$ knockout that diminishes contact inhibition[62] and apoptosis). Though specific drivers and passengers will have differing effects on the birth and death rates, below we will demonstrate that aggregating the effects of mutations into the birth rate, and placing the effects of population size into the death rate, does not alter population dynamics from models where mutational effects are split between the two (figure 3.5 on page 30). Thus, we use:

$$B(n_d, n_p) = \frac{1 + s_d}{1 + s_p} \qquad D(N) = \frac{N}{N^0} \tag{2.1}$$

$$\tag{2.2}$$

where s_d is the fitness advantage (selection coefficient) of a driver, s_p is the fitness disadvantage conferred by a passenger, and N^0 is the initial equilibrium population size—reflecting the effects of the tumor microenvironment. This model assumes 'multiplicative' epistasis and is mathematically equivalent to $B(n_d, n_p) = (1 + s_d)^{n_d}(1 - s'_p)^{n_p}$ when $s'_p = s_p/(1 + s_p)$. In section 3.3 on page 36, we discuss the effects of alternative models of epistasis, specifically an 'additive' epistasis model and a 'two-hit' model of carcinogenesis. Both models exhibit similar behavior to this original model.

For our initial, mathematically-tractable model, the death rate increases linearly

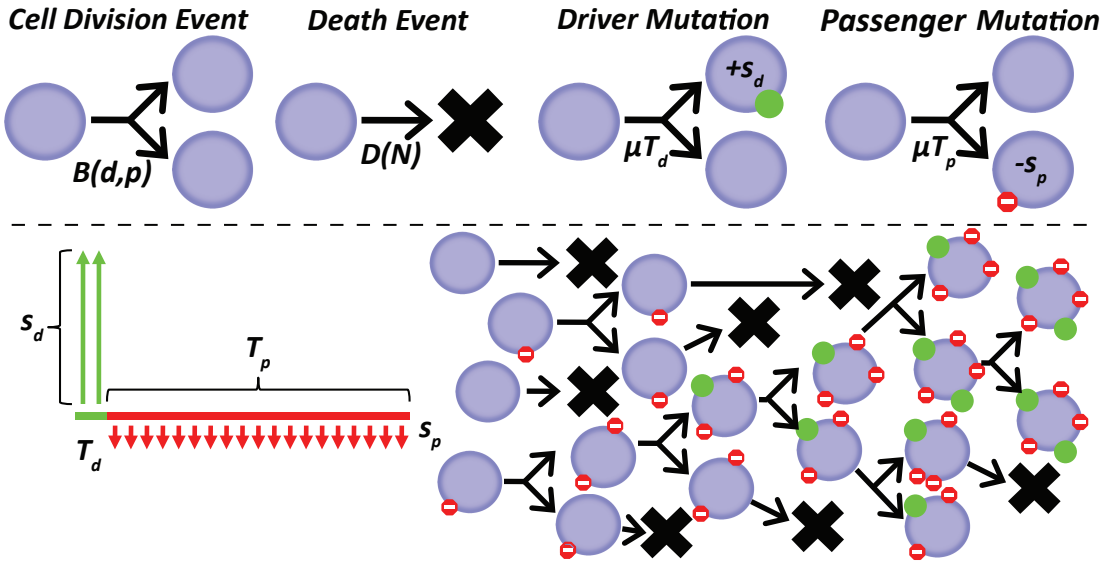


Figure 2.1: A model of cancer progression with drivers and passengers. Our evolutionary model: individual cancer cells can stochastically divide (potentially acquiring new drivers or passengers) and die. A new driver increases the birth rate by s_d, whereas a passenger decreases the birth rate by s_p (equation (2.1) on the previous page). (Lower Left) Drivers arise rarely (reflected by a small T_d), but have a large effect size s_d, while passengers are common (large T_p), but have small individual effects s_p. (Lower Right) A conception of a population stochastically dividing and dying. When a driver arises in the population, its clone expands and carries with it deleterious passengers.

with population size (similar to previous neoplastic[58] and ecological[33] models). However, in later sections, when discussing large tumors (grown to 10^6 cells), we let $D(N) = \text{Log}(1 + N/N^0)$. For small N/N^0, during early carcinogenesis, this reduces to the linear model above, but for large N/N^0 this recapitulates Gompertzian dynamics observed experimentally for large tumors[50]. The death rate's dependence on population size is a coarse approximation of many size-dependent factors that tumors must overcome as they expand via additional drivers: contact inhibition, competition between cells for space and resources (e.g., due to a limited crypt size), homeostatic pressure, hypoxia, angiogenesis, limited paracrine signaling, and immune/inflammatory responses to larger tumors[3].

We do not track individual cell's position in space. Hence, in every model we consider here, the environment must be homogenous and we must ignore the spatial structure of cancer. Previous studies of asexual populations suggest that ignoring spatial structure will (i) underestimate the time for beneficial drivers to sweep through the population and hence the degree of clonal interference, and (ii) overestimate the effective strength of selection, which only acts at the geographic boundary between clones[64,65]. Hence, models considering spatial structure should find that more passengers fixate relative to those that do not, thereby strengthening the conclusions of our model. Nevertheless the effects of spatial constraints have been largely left to future studies and is an area of particular interest to us—especially because it generally increases the amount of genetic diversity in a population.

During division, cells can also acquire driver and passenger mutations. The number of acquired driver mutations during division Δd and the number of acquired passenger mutations during division Δp are Poisson-distributed pseudorandom variables with means μ_d and μ_p, respectively; Thus, for passengers: $P(\Delta p = k|\mu_p) = \frac{(\mu_p)^k e^{-\mu_p}}{k!}$.

These driver and passenger mutation rates are a product of the overall per-locus mutation rate μ multiplied by the respective targeted sizes of drivers T_d and passengers T_p (i.e. number of accessible driver/passenger loci). Hence, $\mu_d = \mu T_d$ and $\mu_p = \mu T_p$.

In this model, mutations arise at a rate proportional to generation time, rather than absolute time. Our choice, therefore, reflects our belief that most mutational processes occur in a cell division-dependent manner (e.g. mutations in cell-cycle checkpoints like p53), rather than independent of cell division. Certainly, some mutational processes occur independent of cell division. However, because the generation time in precancerous cells accelerates only mildly over the course of progression, introducing mutations at an absolute rate would not significantly alter dynamics. Also, mutations that do not arise as a direct result of cell division will often be repaired, unless the cell divides first. Hence, even these mutational processes are proportional to the rate of cell division.

We can now fully define our model in terms of chemical half-reactions in the Gillespie formalism:

$$\begin{aligned} \{n_d, n_p\} &\xrightarrow{B(n_d,n_p)} \{n_d + \Delta d, n_p + \Delta p\} + \{n_d + \Delta d, n_p + \Delta p\} \\ \{n_d, n_p\} &\xrightarrow{D(N)} \emptyset \end{aligned}$$

Notice that population size can vary because of (i) accrued drivers, (ii) accrued passengers, and (iii) stochasticity in the birth & death rates. We can also define a 'generation' in terms of the mean division time:

$$1 \text{ generation} = \frac{1}{1/N \sum_{i=0}^{N} B(n_{d_i}, n_{p_i})}$$

In general we use this measure of time for analysis, as it is a convention in population

genetics.

Many of the above considerations and issues associated with formulating our model, are germaine to all models of tumor progression, not simply the *in silico* models presented here. Consider that recent data on growth rates of human tumors differs from data obtain from mouse models: human tumors grow according to an exponential curve[13], while mouse tumors grow according to a Gompertz curve[30]. Careful mathematical consideration of the differences between a model of progression where growth is exponential, and one where growth is Gompertzian, should allow us to understand when it is necessary to refine the design of simulations and experiments, when these minutia can be ignored, or, instead, when we should alter our conclusions.

2.2 Evolutionary parameters

The evolutionary system above is fully defined by five parameters: s_p, s_d, $\mu_d = \mu T_p$, $\mu_p = \mu T_d$, and N^0. These parameters vary considerably between tumor types (and the mutation rate even varies within tumor types[72]), so we explored a wide range of values centered around literature best-estimates (table 2.1 on page 15). Though driver and passenger alterations take many forms, we first parameterized our model using single-nucleotide substitution data, as these mutations have been more thoroughly quantified. In chapter 5 on page 68, we investigate both alterations and point-mutations and estimate selection coefficients from these data.

Because of extensive cancer heterogeneity and limited quantitative knowledge, we varied all parameters by 2 to 3 orders of magnitude. The ranges we explored centered on values obtained from the literature (table 2.1 on page 15). The mutation rate ($\mu \approx 10^{-8}\text{nt}^{-1} \times \text{division}^{-1}$; range 10^{-10}–10^{-6}) approximates cells with a mu-

tator phenotype[75]. Our initial equilibrium population size ($N^0 \approx 10^3$ cells; range 10^2–10^4) was estimated from hyperplasias within a mouse colonic crypt observed 2 weeks after an initiating APC deletion[24]. The target size for drivers ($T_d \approx 700$ nt; range 70 – 7,000) is approximately 10 potential hotspot mutations per gene (oncogene or tumor suppressor) times 70 driver genes[111]. This value was used in previous simulations[8] and is close to the 571 loci with recurrent mutations in colon cancer[36]. The target size for functional (nonsynonymous) passengers ($T_p \approx 5 \times 10^6$ nt; range 5×10^5–5×10^7) was estimated as 10^3 nonsynonymous loci per gene times 5,000 well-expressed, non-cancer-related genes in cancer[22]. This value is comparable, but less than, a previous estimate of 10 million deleterious loci in cancer[7]; does not attempt to capture the 10^4–10^5 non-coding passenger mutations per cancer genome[101,10]; and yet is thousands of times greater than T_d. The chosen driver strength [$s_d \approx 0.1$ (i.e., 10% growth increase per driver); range 0.01 – 1] was shown to be congruent with cancer onset[8]. Passenger deleteriousness ($s_p \approx 10^{-3}$; range 10^{-1}–10^{-4}) was estimated from to the effects of near-neutral germ-line mutations in humans[34] and randomly introduced mutations in yeast[41].

We consider death to be any process that prevents a cell from replicating indefinitely, i.e. necrosis, apoptosis, senescence, or differentiation. Thus, N represents only cells capable of infinite division and of carrying the (epi)genetic information in cancer. For this reason, our model lacks asymmetric cell divisions, as this yields differentiated cells. We first explore the initial population size N^0 across two orders of magnitude. Hence, our model applies equally well to tumor subtypes dominated by only a small cohort of cancer stem cells and subtypes where cancer may arise from progenitor cells[49]. Later on, we will demonstrate that a population grown to a particular size N is equivalent to a population initialized at N. This path-independence allowing us

Table 2.1: Evolutionary parameters explored in this study. We explored our evolutionary model incorporating driver and passenger mutations across a broad range of parameters. The ranges were motivated by literature estimates discussed previously[83] and in section 2.2 on page 13. Note that in simulations $\mu_d = \mu T_d$ and $\mu_p = \mu T_p$, hence the entire phase space can be explored by only altering μ and T_d/T_p, as altering all three parameters is redundant. In chapter 5 on page 68 we compare our model to epidemiological and genomic data and affirm that these prior published estimates explain the new data well.

Parameter	Symbol	Estimate	Range	Citation
Mutation rate	μ	10^{-8}	10^{-10} - 10^{-7}	75
Driver loci	T_d	700	70 - 7,000	39,111,8
Passenger loci	T_p	$5 \cdot 10^6$	$5 \cdot 10^5$-$5 \cdot 10^7$	7,14
Driver strength	s_d	0.1	0.001 - 1	8,11
Passenger strength	s_p	0.001	10^{-4} - 10^{-1}	41
Initial population size	N^0	1000*	100 -10,000	24

*Estimated from labeled populations in mice colonic crypts 2 weeks after an initiating APC deletion was induced.

to simplify dynamics and explain progression for any initial or intermediate population size.

The most critical constraint of our model is that $T_d \ll T_p$. Without this property a barrier to adaptation is not observed and infinite mutation rates become optimal. This constraint on target sizes for simulations is justified for a number of reasons. A priori, it should be expected that deleterious mutations outnumber advantageous mutations in natural populations simply because natural selection optimizes genomes to their environment—implying that most changes will be neutral or damaging. Indeed, most protein coding mutations and alterations were deleterious or neutral when investigated empirically in fly[51], yeast[117], and bacterial genomes[20]. We consider only moderately deleterious loci here ($s_p \approx 10^{-4} - 10^{-1}$)—which nevertheless account for most nonsynonymous mutations[41,14]. Deleterious mutations outside of this range either do not fixate or negligibly alter progression[83].

There is also considerable evidence that $T_p \gg T_d$ in cancer. As much as 10% of

the human genome is well-conserved and likely deleterious when mutated[61,68]. Conversely, there are only approximately 100-200 potential driver genes[21,39]. If driver loci include only a few specific sites per gene ∼10, then collectively drivers will constitute less than one one-millionth of the genome. Also, accumulated passengers greatly outnumber accumulated drivers (table 1.1 on page 4). This implies that the target size of passengers greatly outnumbers the target size of drivers, as selection can only increase the frequency of advantageous mutations relative to deleterious mutations.

For most of our parameter range $s_d > s_p$, but we also explored exceptions to this rule and captured their dynamics with our analytical model. The selection coefficients of drivers s_d and passengers s_p were estimated from genomics data in the main text and found to be comfortably within the range we explored. Nevertheless, there was good evidence for the range of fitness benefits for drivers before we began our study. A previous study found that an s_d of 0.1 is necessary to obtain waiting times to cancer consistent with age-incidence rates[8]. Later on, we discuss evidence from mice models that support this approximate value of s_d.

The mutation rate not only varies considerably between tumors[72], but also varies as a tumor evolves. We believe that genomic instability happens early during progression, as this has been shown experimentally some tumors[109] and suggested to be the first event during progression by virtue of theoretical considerations[96]. However this presumably differs from tumor to tumor. Thus, by developing a theoretical understanding of the process, we hope to gain some intuitive understanding of how variation in mutation rate over time might alter dynamics. For example, we show later that for tumors far below the critical mutation rate μ^*, variation in the mutation rate increases or decreases the rate of accumulation of drivers and passengers equally. However, near or above μ^*, variation in the mutation rate has a profound impact on

driver's probability of fixation and effect size.

2.3 Previous evolutionary models of cancer progression

Two previous investigations of cancer progression have considered deleterious passengers and found that they have a minimal impact on progression[7,104]. The first paper assumes that passengers are effectively lethal to cancer cells (i.e. $s_p \to \infty$). They conclude that deleterious passengers are unimportant because they are quickly weeded out of the population. This is consistent with our results (see chapter 3 on page 22, or figure 4.1 on page 46), however we believe that our best-estimate of $s_p \approx 10^{-3}$ is more reasonable. Our justification for this is discussed above, but also supported by a lack of evidence for negative selection in cancer genomes (see figure 5.2 on page 75), suggesting that deleterious passengers must be accumulating and cannot be lethal. Indeed, we observe the accumulation of passenger mutations within housekeeping genes and specifically at loci predicted to be deleterious by classification algorithms (see figure 5.1 on page 72). To us, this is clear evidence that s_p cannot be large enough to prevent the fixation and accumulation of most deleterious passengers.

In the second study that models deleterious passengers as part of an inquiry on the effects of genomic instability[104], the authors assume that there exist only ∼100 housekeeping genes in cancer that are deleterious when mutated. Again, we observe similar behavior in our simulations: when $T_p \approx T_d$, passengers do not appreciably alter progression. However, our best-estimate of the number of relevant deleterious genes is 50× larger than their estimate (table 2.1 on page 15), while the paper discussed in the preceding paragraph argues that T_p is *even larger* than our estimate (100× greater than the estimate in this second paper[7]). There are two reasons for this discrepancy

in parameter choice. First, the article considers only deleterious housekeeping genes, while we believe many other genes and non-coding sequences (e.g. regulatory DNA sequences or microRNAs) could potentially be deleterious to cancer cells[104]. Secondly, we find that the number of reported housekeeping genes is much larger than 100 (3,804 genes are classified as 'housekeeping' in[31], which also reviews other, similar estimates). We believe most genes could be deleterious when mutated because (i) it has been proposed that passengers might invoke an immune reaction to tumor cells[120], and (ii) because they may cause cytotoxicity via protein disbalance and aggregation[108]. This later mode of damage should be applicable to nearly all expressed genes within the tumor, which constitute more than half of the 26,588 identified genes in the human genome[31].

Lastly, the article[104] remarks that passenger's drag on population size weakens as s_p increases beyond 0.01. In our theoretical work below, we show that the optimum s_p for slowing progression s_p^* is approximately proportional to the mutation rate of passengers ($s_p^* \sim \mu_p$). Hence, for the authors chosen target size of passengers, their finding is correct; however, given our larger estimate of μ_p (discussed above), increases of s_p beyond 0.01 continue to increase the drag of passengers on cancer progression.

Because the relevance of deleterious passengers depends upon evolutionary parameters, their effects may be neglected in certain circumstances. In figure 4.1 on page 46, we identify the evolutionary regimes where passengers dominate, where they compete with drivers, and where they can be probably be neglected. We believe and present evidence that passengers are relevant for progression in most carcinomas, but we also observe that lymphomas fixate very few passengers (table 1.1 on page 4). All these concerns underscore the importance of further investigating the evolutionary parameters of cancer progression for various tumor types to ascertain passengers relevance to

cancer progression.

2.4 Where to begin?

Before describing the entire dynamics of our model, it is useful to consider the difference between our simulations initiated at their stationary size (N^0 cells) and simulations initiated at 1 cell. In the absence of mutations, an initial population of one cell will grow logistically until it reaches the stationary size. Hence, it takes approximately $\text{Log}_2[N^0] \sim \text{Log}_2[10^3] \sim 10$ generations for the initial cell to approach stationary size. This is far shorter than the average time required for cancer progression ($\sim 10,000$ generations) and the time required for a new driver to accumulate ($\sim 1/(\mu_d N^0 s_d) \sim 1,000$ generations). Thus, our choice of initiating a tumor at one cell versus N^0 does not significantly alter the conclusions of our model.

This comparison of timescales also suggests that cancers are almost always near their stationary size:

$$\overline{B(n_d, n_p)} \approx D(N)$$

We test this conclusion in simulations and found that it is a excellent approximation of tumor size (figure 3.9 on page 37). If we assume $\overline{B(n_d, n_p)} \approx B(\overline{n_d}, \overline{n_p})$, then a relationship between the number of drivers and passengers in a tumor and its size is

obtainable:

$$
\begin{aligned}
B(\overline{n_d}, \overline{n_p}) &\approx D(N) \\
\frac{(1+s_d)^{\overline{n_d}}}{(1+s_p)^{\overline{n_p}}} &\approx \mathrm{Log}[1 + \frac{N}{(e-1)N^0}] \\
\overline{n_d}\mathrm{Log}(1+s_d) - \overline{n_p}\mathrm{Log}(1+s_p) &\approx \mathrm{Log}[\mathrm{Log}[\frac{N}{N^0}]]
\end{aligned}
$$

$$
\overline{n_d}s_d - \overline{n_p}s_p \approx \mathrm{Log}[\mathrm{Log}[\frac{N}{N^0}]] : s_d, s_p \ll 1 \tag{2.3}
$$

This final equation suggests that there exist a linear relationship between drivers and passengers among tumors with similar s_d and s_p, which we assume is the case for tumors of the same tissue of origin. The relationship should be relatively robust to tumor size, but sensitive to the fitness effects of drivers and passengers. Moreover, changes in the functional form of $D(N)$ will alter the y-intercept of this linear relationship, but not the slope of the relationship. Hence, we can draw conclusions about the relative strength of drivers versus passengers (s_d/s_p) without knowing the exact constraints on population size. We tested and verified this prediction of a linear relationship between drivers and passengers in chapter 5 on page 68.

3

Deleterious passengers accumulate & alter progression

3.1 Moderately Deleterious Passengers Fixate and Alter Cancer Progression.

Figure 3.1 on the next page presents typical population trajectories of cancer beginning at the first driver mutation. All trajectories consist of intervals of rapid growth and gradual decline. A new driver leads to a clonal expansion of the subpopulation carrying this driver, causing short periods of rapid growth. Growth stops when the effect of this driver is balanced by the death rate, which increases with population size. While the population waits for the next driver to arise, passengers steadily accumulate, causing a gradual decline of population size. Together, these processes cause trajectories to grow in a sawtooth pattern. Simulated tumors exhibit either unconstrained growth or regression, often after a period of dormancy (figure 3.1 on the following page).

The phenomena of dormancy and spontaneous regression, observed both in our model and clinically[2], do not occur in models lacking deleterious passengers.

Importantly, simulations show that hyperplasias that progress to clinical size (i.e., 10^6 cells) accumulate many deleterious passengers. This evasion of purifying selection and fixation of deleterious passengers is an unexpected result not programmed into

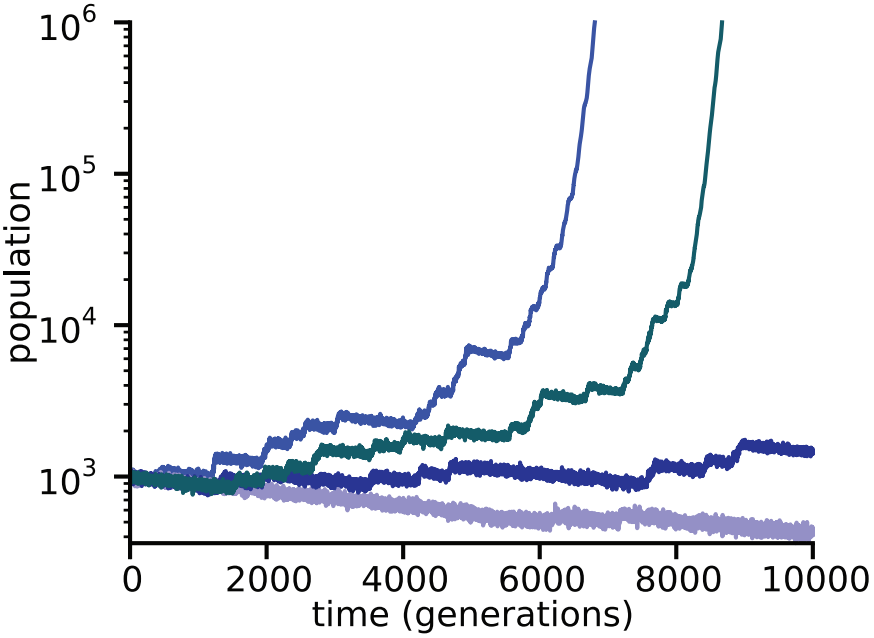


Figure 3.1: Precancerous lesions can remain dormant, rapidly grow to cancer, or regress. Despite identical parameters, simulated cancer progression trajectories exhibit markedly different behavior. Some populations regress to extinction or exhibit long periods of dormancy. The accumulation of strongly beneficial drivers alongside moderately deleterious passengers results in a reversed-sawtoothed pattern of growth.

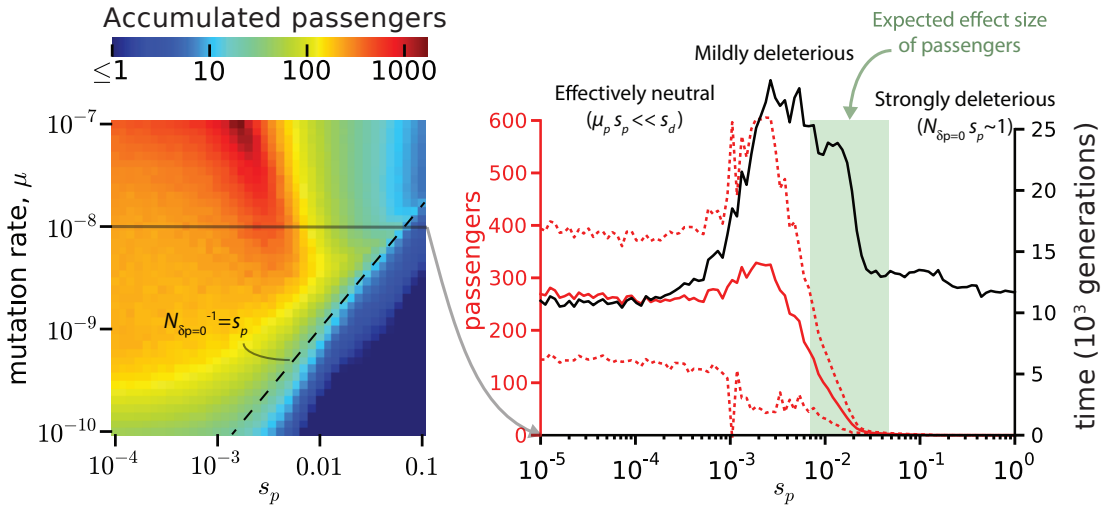


Figure 3.2: Passengers accumulate across a broad range of mutation rates and passenger deleteriousness. (Left) The number of accumulated passengers increases with mutation rate and depends, non-monotonically, on passenger strength. Passengers accumulate exceedingly slowly once the population size of the fittest class exceeds the inverse strength of passengers. (Right) Moderately deleterious passengers alter cancer progression and mostly evade selection. Passengers of intermediate fitness effect s_p prolong the time to cancer and accumulate in large, highly variable quantities (mean in solid red; ± 1 S.D. in dashed red). Moderately deleterious passengers affect cancer only if they are strong or frequent enough to be comparable to the effects of drivers, yet weak enough to avoid selection. Experimentally observed fitness effects of random point mutations in YFP in yeast ranged from 0.007 to 0.028 (green shading)[41].

the model. Although the exact number of accumulated passengers depends on μ and s_p (figure 3.2), $10^2 - 10^3$ deleterious passengers are obtained for a broad range of parameters, consistent with the numbers of nonsynonymous substitutions observed in cancer genomics studies (table 1.1 on page 4), suggesting that observed passengers in sequencing data can be moderately deleterious.

Figure 3.2 shows that moderately deleterious, rather than highly deleterious or neutral, passengers have a major effect on cancer progression. Indeed, almost-neutral passengers have very little effect on cancer cells, and passengers of large effect do not accumulate[7]. By slowing progression to cancer, moderately deleterious passengers

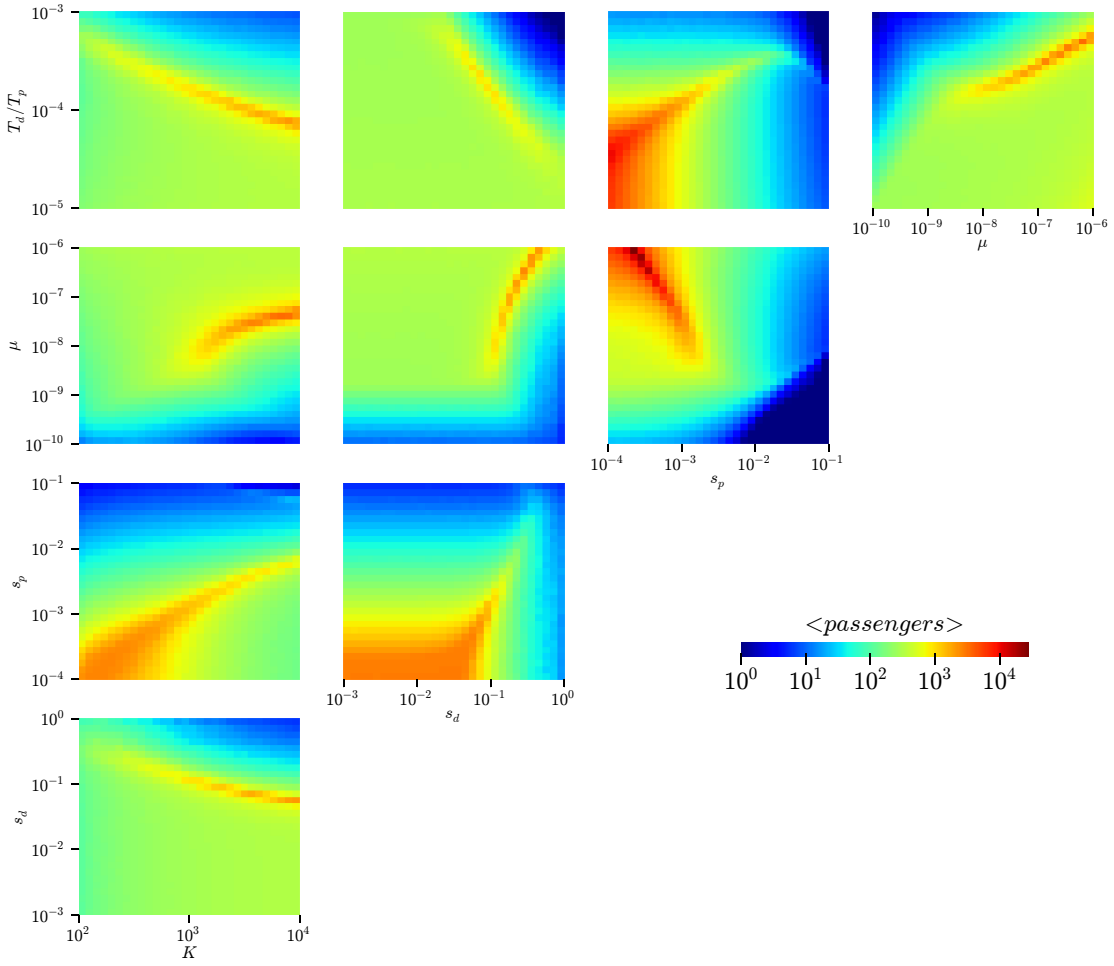


Figure 3.3: Passengers accumulate non-monotonically across the entire parameter space. We simulated tumor progression across all parameters. For each heat map, the three parameters not shown on the x or y axes were set to estimated values (table 2.1 on page 15). For each element in the heat map, 200 trajectories were simulated until they progressed to cancer, progress to extinction, or until 15,000 generations (50 years), whichever was sooner. We define progression as doubling of cancer size because, after doubling, populations progress quickly to cancer and acquire few additional passengers. The mean number of accrued passengers is shown with color. Because variation in the mutation rate and s_p alter the time to cancer and the probability of progression to cancer, the number of passengers depends non-monotonically on both parameters. Behavior across this phase space will be described in greater detail in (chapter 4 on page 41).

accumulate in even greater numbers than neutral mutations despite their slower accumulation rate (figure 3.2 on page 24). Importantly, we find that moderately deleterious passengers affect progression for s_p from 10^{-3} to 2×10^{-2}, which subsumes the best literature estimates of the strength of deleterious mutations[41,34]. Such small selection coefficients for individual passengers are typically undetectable in cell cultures, yet critical for long-term cancer dynamics.

3.1.1 Passengers accumulate via genetic draft and Muller's Ratchet

We then studied how deleterious mutations can accumulate despite negative selection. Previous studies have calculated the rate of accumulation of deleterious mutations in the absence of clonal expansions[16,47,92]. In figure 3.4 on page 28, we identify two previously known processes that allow passengers to evade negative selection in cancer: hitchhiking alongside a driver and Muller's ratchet[33]. Deleterious passengers hitchhike when the cell they reside in acquires a new driver, which then leads to a clonal expansion and fixation of all the mutations in that cell. Muller's ratchet, in turn, is a process of gradual accumulation of deleterious mutations and population decline in the absence of drivers. In Muller's ratchet, a mutation-selection balance arises after driver sweeps, which creates a steady-state Poisson distribution of the number of passengers per cell with mean and variance μ_p/s_p (first described in ref.[53] and discussed further in chapter 4 on page 41). The fittest subpopulation—cells with the fewest passengers: $N_{\delta p=0} \sim Ne^{-\mu_p/s_p}$ cells—is much smaller than the whole population, so it can spontaneously shrink to extinction (figure 3.4 on page 28). When back mutations are rare, such an extinction leads to the irreversible loss of this least-mutated fraction of cell and corresponds to a 'click' of Muller's ratchet[33]. This process is especially rapid during clonal expansions when the size of the expanding clone is small. Both of the above

processes, well known in population genetics, are augmented in cancer because of the presence of strong drivers.

3.1.2 Various distributions of passenger strengths exhibit similar behavior to fixed-effect model

We then relaxed our assumption that s_d and s_p are constant for all passengers, by simulating cancer progression with drivers/passengers drawn from distributions of effect sizes (figure 3.5 on page 30). There does appear to be some greater variation in the waiting time to cancer when drivers drawn from a distribution of effect sizes are used. However, the overall dynamics appear qualitatively similar, and suggest that our model is a good approximation of more sophisticated models.

Like drivers, we simulated trajectories with passenger mutations drawn from several potential fitness distributions[94]. The effect of each passenger x, was drawn from either an exponential distribution, a Log-Normal distribution, or a Gamma distribution with the following density functions.

$$\begin{aligned}
\text{Exponential}(\text{x}|\text{s}_\text{p}) &= \text{Exp}[-\text{x}/\text{s}_\text{p}]\text{s}_\text{p}^{-1} \\
\mathscr{LN}(x|s_p) &= \frac{1}{x\sqrt{2\pi}}\text{Exp}[-\frac{(\text{Log}(\text{x}/\text{s}_\text{p}) + \frac{1}{2})^2}{2}] \\
\Gamma(x|k=2, s_p) &= s_p^{-2} x \text{Exp}[-2\text{x}/\text{s}_\text{p}]
\end{aligned}$$

All distributions have a mean value of s_p of 0.001.

The strength of passenger mutations affects their fixation probability. For passengers, the variation in fitness within a population is mostly invariant to the type of dis-

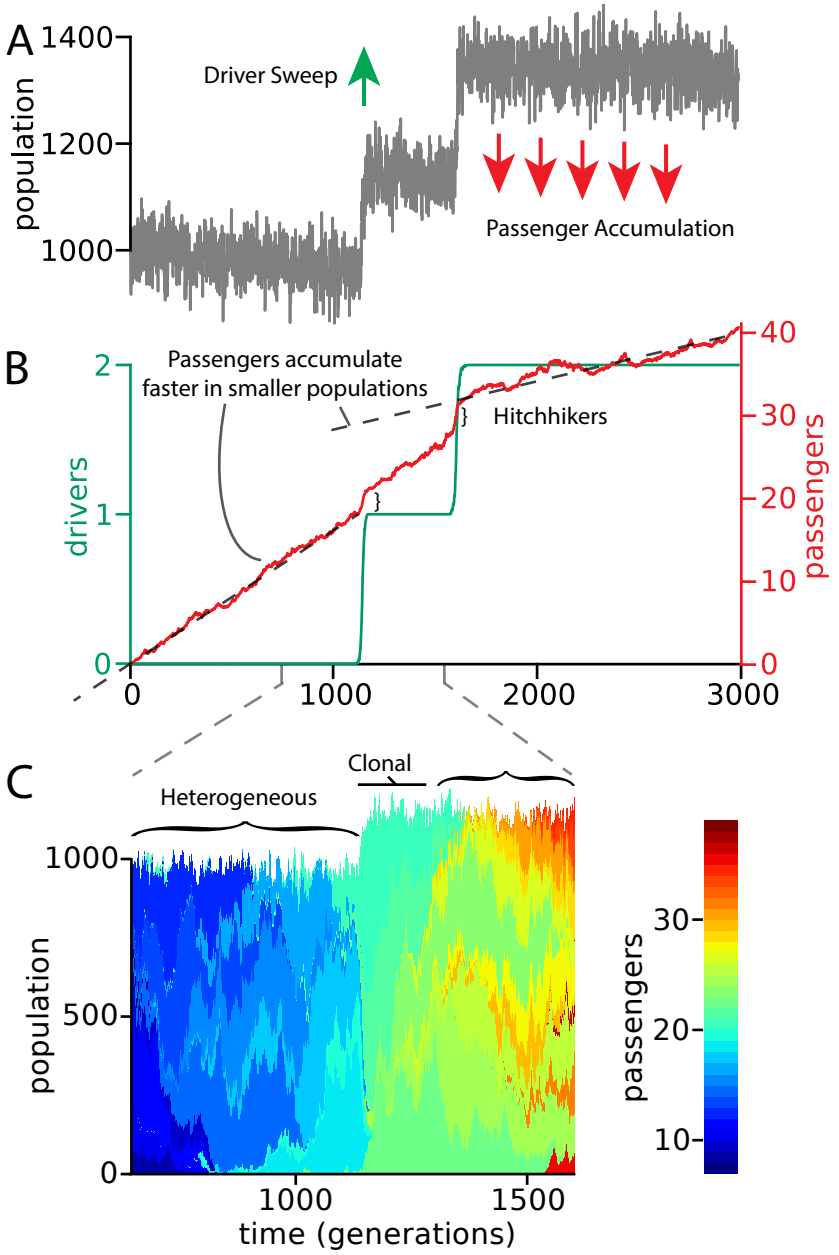


Figure 3.4: Mechanism of passenger accumulation. **(A)** Spurts of population growth, caused by the acquisition of a new driver, are interspersed with a gradual decline due to passenger accumulation. **(B)** Passengers accumulate both steadily between the arrival of drivers and by hitchhiking during clonal expansions. **(C)** Each sub-clone, containing a unique number of passengers (shown by color), grows and declines stochastically, eventually to extinction. In between drivers, the population becomes heterogeneous. A new driver will promotes only one clone, creating a clonal population. Afterward, new mutations on top of the previous hitchhikers restore heterogeneity.

tribution of passenger effects (figure 3.6 on page 31). Negative selection against passenger fixation appears to be largely inefficient, except for highly deleterious passengers (figure 3.5 on the following page). The significant variance in cell fitness within the population, caused by deleterious passengers (figure 3.6 on page 31), also affects the probability of driver fixation. Because a driver will generally occur in a cell of average fitness, it is unlikely to fixate unless its new fitness is greater than the fittest cells. The difference between the fittest cells and average cells in the population is approximately μ_p and independent of s_p (figure 3.6 on page 31)[44]; therefore, a driver must confer a benefit greater than μ_p to fixate (discussed in detail in section 4.2 on page 63). This argues that weak drivers are unlikely to fixate in cancer or be observed in genomic sequencing.

The interplay between the deleterious mutation rate and variance in population fitness suggests that precancerous populations with larger mutation rates will exhibit suppressed levels of negative selection. Indeed, this appears to be the case for exponentially-distributed passenger effect sizes (figure 3.6 on page 31). This finding may be testable in future genomic studies by investigating the degree of purifying selection in sequenced cancer genomes and comparing this value to the cancer's overall mutation rate and each mutation's expected deleteriousness.

In section 4.3 on page 65, we discuss an analytical model of passengers with varying effect size that is inspired by the results in this section.

In summary, our simulations demonstrate that despite the moderately deleterious effect of individual passengers, they accumulate in large numbers during neoplastic progression, reducing the fitness of cancer cells and altering the course of neoplastic progression. We find several reasons why deleterious passengers accumulate more than might be expected a priori: (i) mutator phenotypes [a hallmark of cancer[75]] ac-

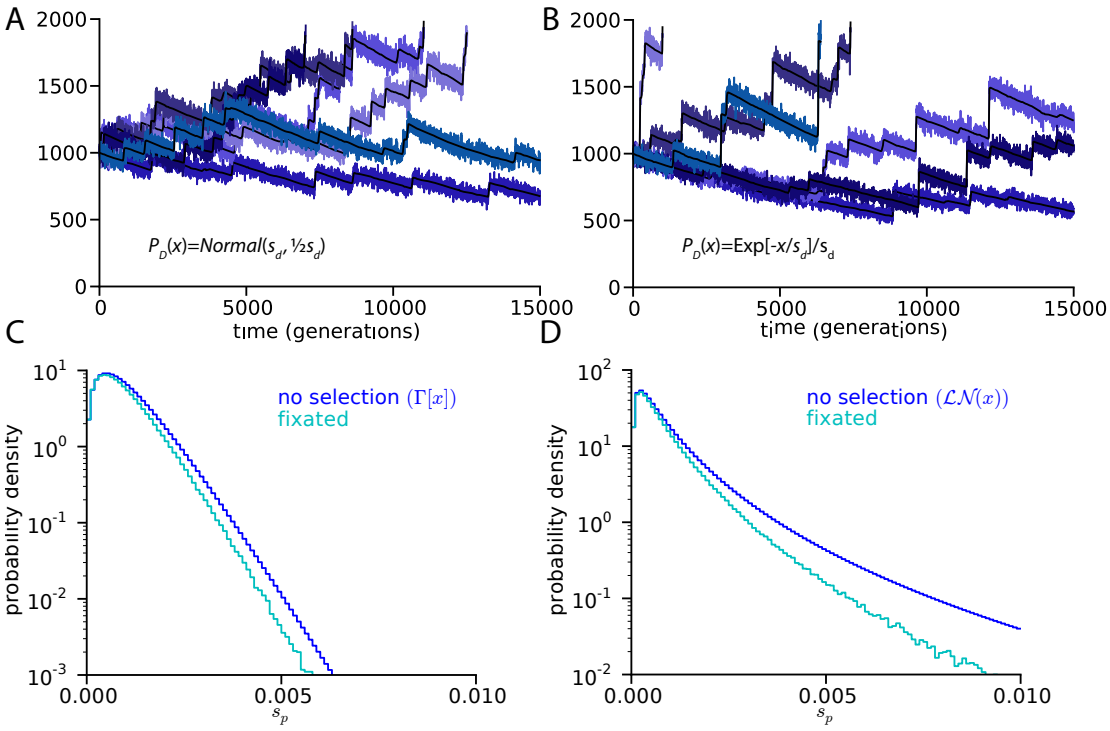


Figure 3.5: Drivers and passengers of varying effect size behave qualitatively similar to mutations of fixed effect size. **(A)** Six trajectories where driver loci are assigned a fitness advantage by sampling from a Gaussian distribution or **(B)** an Exponential distribution appear qualitatively similar to trajectories with fixed fitness effects (figure 3.1 on page 23). P_D represents the probability density function from which drivers were drawn. For the Normal distribution $P_D(x) = \sqrt{\frac{2}{s_d \pi}} \mathrm{Exp}[-\frac{2(x-s_d)}{s_d^2}]$, while $P_D = \mathrm{Exp}[-x/s_d] s_d^{-1}$ for the exponential distribution. All trajectories have a mean s_d of 0.1. Likewise, **(C)** passengers drawn from a Gamma distribution or **(D)** Log-normal distribution accumulate at a similar rate to passengers with fixed effects, however passengers with very strong effect sizes are weeded out.

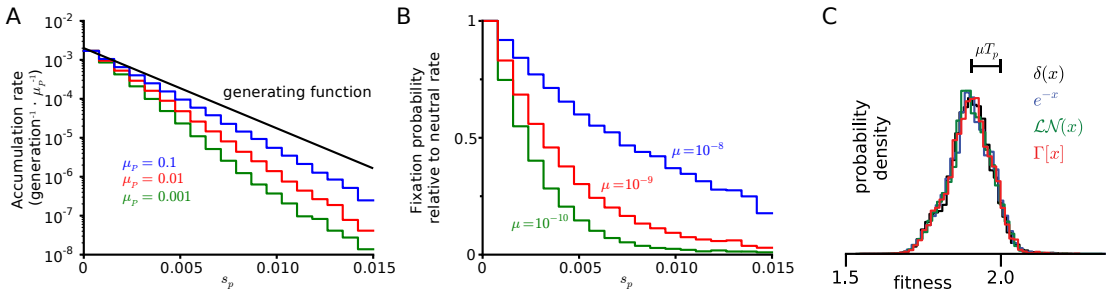


Figure 3.6: Mutation rate of deleterious passengers determines properties of passengers with varying effect size. **(A)** Cancer was modeled with deleterious passengers drawn from an exponential distribution during cell division with rate μ_p. Passengers with very weak deleteriousness are both more frequent in the distribution and fixate more often. **(B)** Weak passengers fixate at a nearly neutral rate, while strong passengers can be weeded-out by natural selection, however this becomes less likely as the mutation rate increases. **(C)** Population dynamics do not change noticeably when passengers were drawn from various distributions of fitness distributions (described above). The variance in fitness is proportional to the mutation rate and much larger than the mean effect size of passengers (i.e. $\mu_p > \bar{s_p}$). Thus, most passengers segregate in the population for an extended period of time, affect dynamics, and can fixate via genetic draft in addition to Muller's Ratchet.

celerate accumulation rates; (ii) small population sizes in the early stages of cancer progression enhance accumulation rates; (iii) driver-induced bottlenecks and hitchhiking contribute additional passengers; (iv) passengers prolong progression—offering more time for accumulation; and (v) passengers arising as part of a distribution of deleteriousness fixate more often than equivalent passengers considered in isolation. These first three phenomenon, though undocumented in cancer theory, have been previously observed in population genetics[78].

3.2 Passengers can prevent cancer progression when tumors are below a critical population size

3.2.1 A critical population size

Figure 3.8 on page 35 shows the dynamics $N(t)$ of individual populations starting at different initial sizes N^0, which correspond to different potential hyperplasia sizes (trajectories begin immediately after a stem cell acquires its first driver, see section 2.4 on page 19 for a discussion of dynamics before this point). Populations exhibit two ultimate outcomes: growth to a macroscopic size (i.e. cancer progression), or extinction, which depend on a critical population size N^*. Larger populations ($N > N^*$) generally commit to rapid growth, while smaller populations ($N < N^*$) generally commit to extinction.

To understand the origin of this critical population size N^*, we examined the short-term dynamics of populations. All trajectories follow a reversed saw-toothed pattern (figure 3.8 on page 35), resulting from a tug-of-war between drivers and passengers. When a new driver arises and fixates in the population, the population size increases to a new stationary value. In between these rare driver events, the population size gradually declines due to passenger accumulation. The relative rates of these competing processes determine whether a population grows rapidly or goes extinct.

The value of N^* can be identified by considering the average change in population size over time ($\langle dN/dt \rangle$), which is the average population growth due to accumulating drivers (v_d) minus the population decline due to accumulating passengers (v_p). When a new driver fixates, the population size immediately increases by $\Delta N = N s_d$. These jumps occur randomly at a nearly constant rate $f = \mu_d N s_d$ (a product of the driver

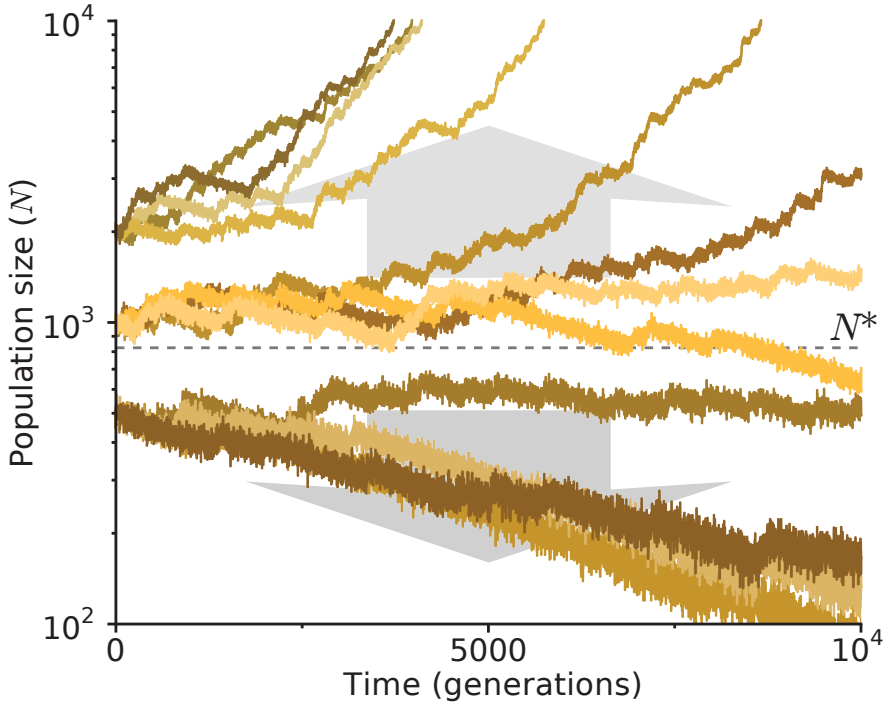


Figure 3.7: Tug-of-war between drivers and passengers leads to a critical population size. Population size verses time of simulations initiated at various sizes ($N^0 = 500, 1000, \text{or } 2000$). Larger population commit to rapid growth, while smaller populations commit to extinction, and intermediate populations bifurcate. A critical population size N^* appears to separate the two regimes.

occurrence rate $\mu_d N$, and its probability of fixation $s_d/(1 + s_d) \approx s_d$). Thus, the velocity due to drivers is $v_d = f\Delta N = \mu_d N^2 s_d^2$. Similarly, passengers' velocity $v_p = \mu_p N s_p$ is a product of their rate of occurrence $\mu_p N$; effect on population size $N s_p$; and fixation probability ($\sim 1/N$ when passengers are near-neutral, see chapter 4 on page 41 for more precise estimates). Thus, we obtain:

$$\left\langle \frac{dN}{dt} \right\rangle = \mu_p s_p N \left(\frac{N}{N^*} - 1 \right) \tag{3.1}$$

$$N^* = \frac{T_p s_p}{T_d s_d^2} \tag{3.2}$$

Here, N^* is the critical population size. In this equation, a population's mean velocity is negative below N^* and positive above N^*, explaining why populations above N^* grow rapidly and populations below N^* continually decline.

These dynamics suggests that N^* constitutes an effective barrier to cancer, which can prevent most cancers from progressing (figure 3.8 on the following page) . Simulations support this conclusion, as they exhibit a sharp transition in the probability of progression at N^* (figure 3.8 on the next page). Indeed, drastically different probability curves collapse onto a single curve when N^0 is rescaled by N^*.

By considering only the average dynamics, we miss the variability of outcomes in adapting populations. In the next chapter, we formulate and solve a stochastic generalization of equation (3.1) that explains the variability of outcomes well. Outcome variability depends only upon N^* and the strength of drivers s_d. Higher values of s_d lead to larger stochastic jumps ΔN, which increases variability and leads to more gradual changes in the probability of cancer across N^0.

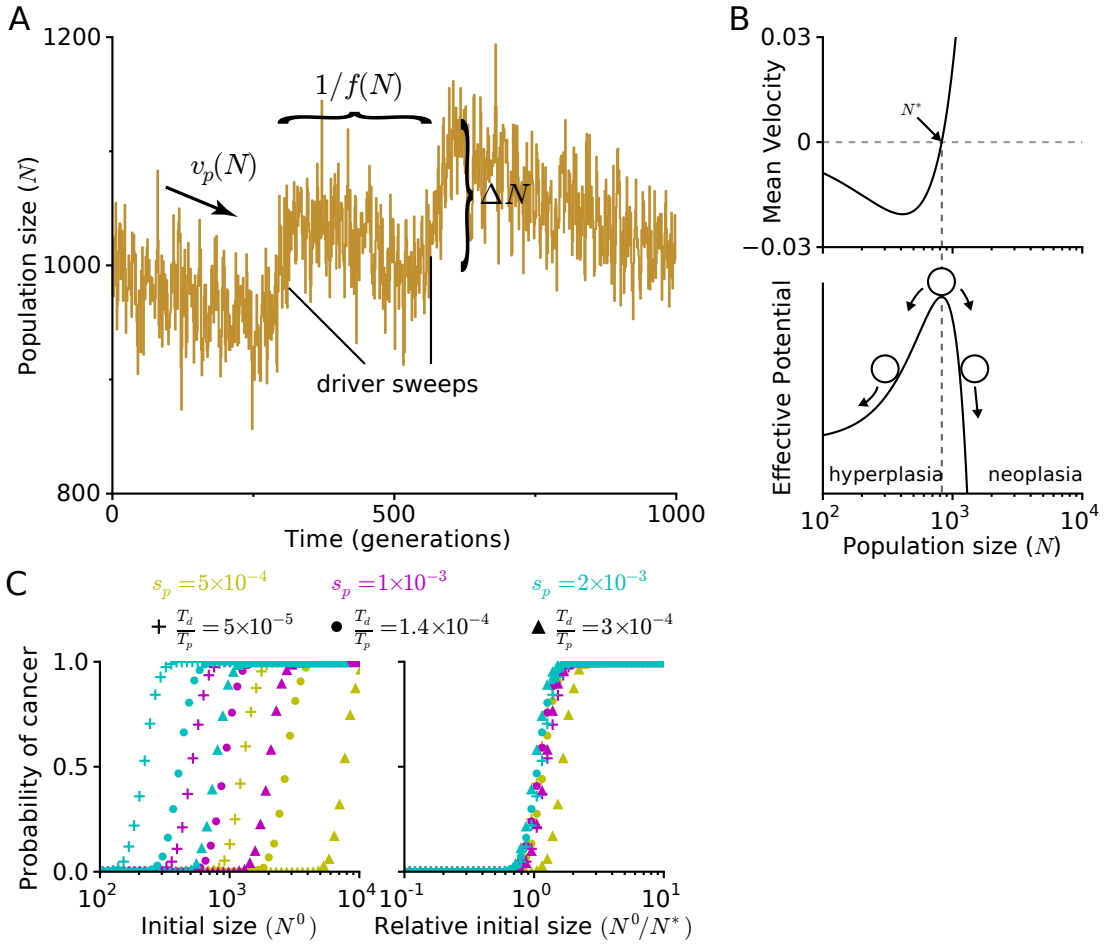


Figure 3.8: Population dynamics can be explained by a 1-D random walk. **(A)** A segment of a trajectory from shows periods of rapid growth and gradual decline. New drivers arrive with a frequency $f(N)$, abruptly increasing the population size by an amount ΔN. Passenger accumulation causes a gradual decline with rate v_p. **(B)** Analytically computed (equation (3.1) on the previous page) mean velocity of population growth (top) and an effective barrier (bottom) as a function of population size N. Velocity is negative below N^* and positive above it. **(C)** The probability of adaptation (cancer) as a function of initial population size N (left) and a relative initial population size (N/N^*, right) for nine sets of evolutionary parameters. Curve collapse is observed, suggesting that simulations behave similarly when plotted relative to N^*.

3.3 Effects of various models of epistasis

Here, we consider two alternative models of epistasis between mutations: a model where mutations interactive additively (rather than multiplicatively), and a 'two-hit' model where the first driver mutation confers no fitness benefit. This later model is meant to approximate certain tumors, particularly blood cancers, where initial proliferation-rate-increasing mutations are only beneficial once apoptotic barriers are eliminated via further driver mutations.

In our additive model, we assume that the fitness benefit and costs of drivers and passengers is $s_d n_d - s_p n_p$. This model lacks the symmetric properties of mutations in our multiplicative model and defies the commonly-used definition of absolute fitness for a mutation: $s_{\text{absolute}} = \frac{N_{\text{after}}}{N_{\text{before}}}$. However, figure 3.9 on the next page demonstrates that it behaves qualitatively similar to our original model.

3.3.1 A two-hit model of cancer progression

It has been proposed that driver mutations may only be beneficial in a certain genetic context[6]. Oncogenes like c-Myc and k-Ras have been shown to induce senescence in some cancer cell lines unless they are accompanied by mutations in p53 or other associated proteins[32]. Likewise, many tumor suppressors mutations are recessive and require a second 'Loss of Heterozygosity' (LOH) event to impart their phenotypic effects (although many tumor suppressors are at least partially haploinsufficient). These types of mutations can be thought of as operating via a two-hit processes: the first driver event confers no change to cell fitness, while the second genetic event confers the benefit of both mutations.

Two-hit models have been studied previously in cancer evolution (Refs.[58,88,96] to

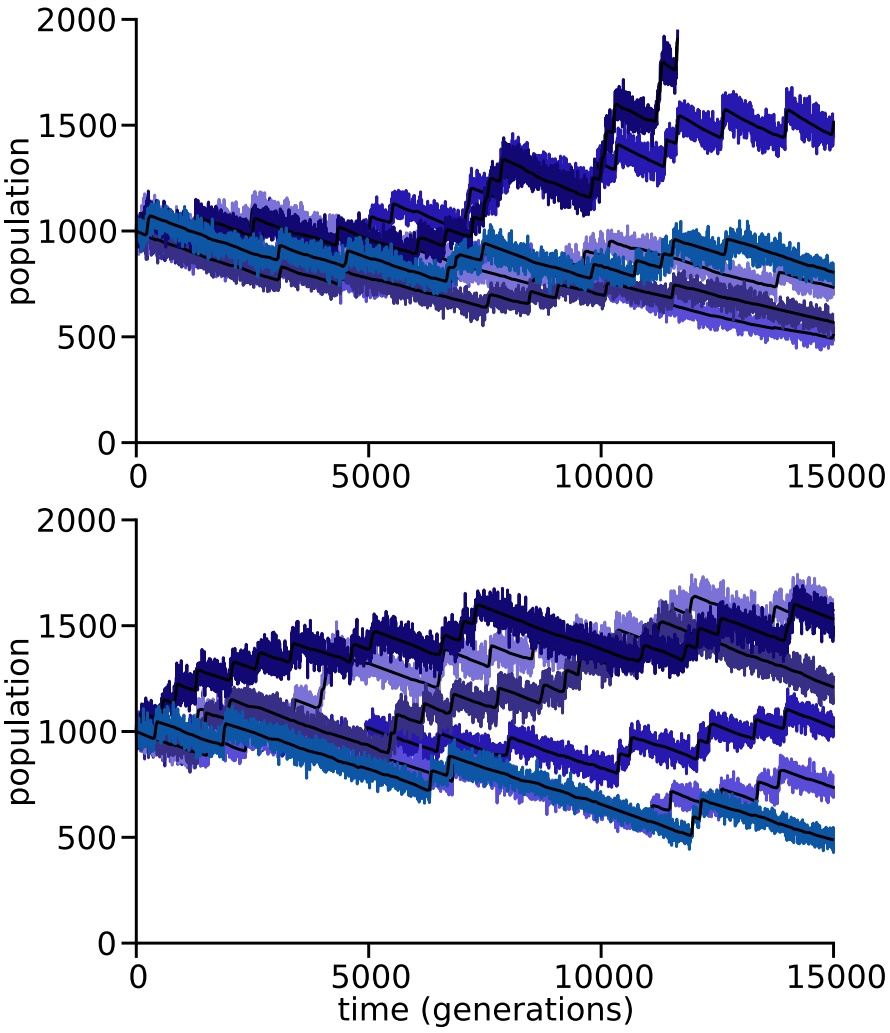


Figure 3.9: Additive and multiplicative epistasis behave similarly. **(A)** Six trajectories (blue shades) with rates $B(n_d, n_p) = \frac{(1+s_d)^{n_d}}{(1+s_p)^{n_p}}$ and $D(N) = \frac{N}{N^0}$ appear qualitatively similar to **(B)** simulations with rate $B(n_d, n_p) = 1 + s_d n_d - s_p n_p$ and an identical death rate. They are also mathematically equivalent to first order expansion of (n_d, n_p). Black lines represent N which satisfies $D(N) = B(n_d, n_p)$. The strong overlap of this black line with the observed population size indicates that birth and death are balanced throughout progression. For this reason, a model where mutations alter death rates, rather than birth rates, would be equivalent to our model provided that time is measured in units of generations.

name a few), but never in the presence of frequent moderately-deleterious passenger mutations. For this reason, we considered our originally defined model above, modified such that the first driver mutation confers no benefit to the cell, while the second driver mutation confers the benefit of both mutations, and all remaining mutations confer a benefit of s_d. Hence,

$$w(n_d, n_p) = \begin{cases} (1+s_p)^{n_p}, & \text{if } n_d < 2 \\ (1+s_d)^{n_d}(1+s_p)^{n_p}, & \text{otherwise} \end{cases}$$

We kept all other properties of the model the same and investigated the result of this permutation in figure 3.10 on page 40. We observe three changes: (1) while the critical barrier to progression remains, its location N^* increases; (2) trajectories slowly decay in a long period of stasis that allows additional passengers to accrue and delays progression; (3) the transition from the non-adaptive to adaptive regime is slower, as the period of stasis is highly variable. We can understand these observations by first considering the mean time of stasis (time until the second driver mutation fixates) for the population. The probability that a second driver fixates in the population at time t is simply the probability that a cell already harbors the first driver mutation ($\mu_d t$), times the total number of cells in the population ($N_0 e^{-v_p t}$), times the probability that the second driver arises and sweeps through the population ($\mu_d \frac{2s_d}{1+2s_d}$). Hence, the probability that stasis will end at generation t is $P_{\text{exit}}(t) = N_0 \mu_d^2 \frac{2s_d}{1+2s_d} t e^{-v_p t}$. Thus, the probability that stasis last a certain number of generations $P_{\text{stasis}}(t)$ is the proba-

bility of not exiting in the prior generations and also exiting at generation t:

$$\begin{aligned} dP_{\text{stasis}}(t) &= -P_{\text{exit}}(t)dtP_{\text{stasis}}(t) \\ P_{\text{stasis}}(t) &= e^{-\int_0^t P_{\text{exit}}(t')dt'} \end{aligned} \tag{3.3}$$

From simulations (figure 3.10 on the next page), it is clear that most simulations that exit this early stasis period go onto progress to cancer. Hence to a first approximation, the new $N^{*,\text{two-hit}}$ is the value of N_0, where $P_{\text{stasis}} = 50\%$, thus:

$$\begin{aligned} \text{Log}(2) &= \int_0^\infty \text{N}^{*,\text{two-hit}} \mu_\text{d}^2 \frac{2\text{s}_\text{d}}{1+2\text{s}_\text{d}} \text{te}^{-\text{v}_\text{p}\text{t}} \text{dt} \\ N^{*,\text{two-hit}} &= \frac{\text{Log}(2)(1+\text{s}_\text{d})\text{v}_\text{p}^2}{2\mu_d^2 s_d} \end{aligned}$$

This predicts $N^{*,\text{two-hit}}$ to be 883 for the trajectories plotted, which is within a factor of two from the 50% success point. It is somewhat below the observed tradition point presumably because (i) trajectories that exit this stasis period far below N^* never progress, and (ii) segregating passengers may interfere with the first beneficial driver. Collectively, these results suggest that our mathematical framework of our model is generally applicable to cancers where the first driver isn't beneficial, but that there are also some dramatic differences that warrant further investigation.

While two-hit models of progression are most likely applicable in many tumors, we do not believe that these models are more universal than our original formulation. Consider that in experiments directly measuring the change in cell fitness upon activating mutations in k-Ras, an immediate increase in proliferation was observed[119]. Also, many tumor suppressors are haploinsufficient (e.g. Dicer[69], p27[100], CDC4[80], p18[99]). These results suggest that driver mutations are often advantageous, after only one mutation.

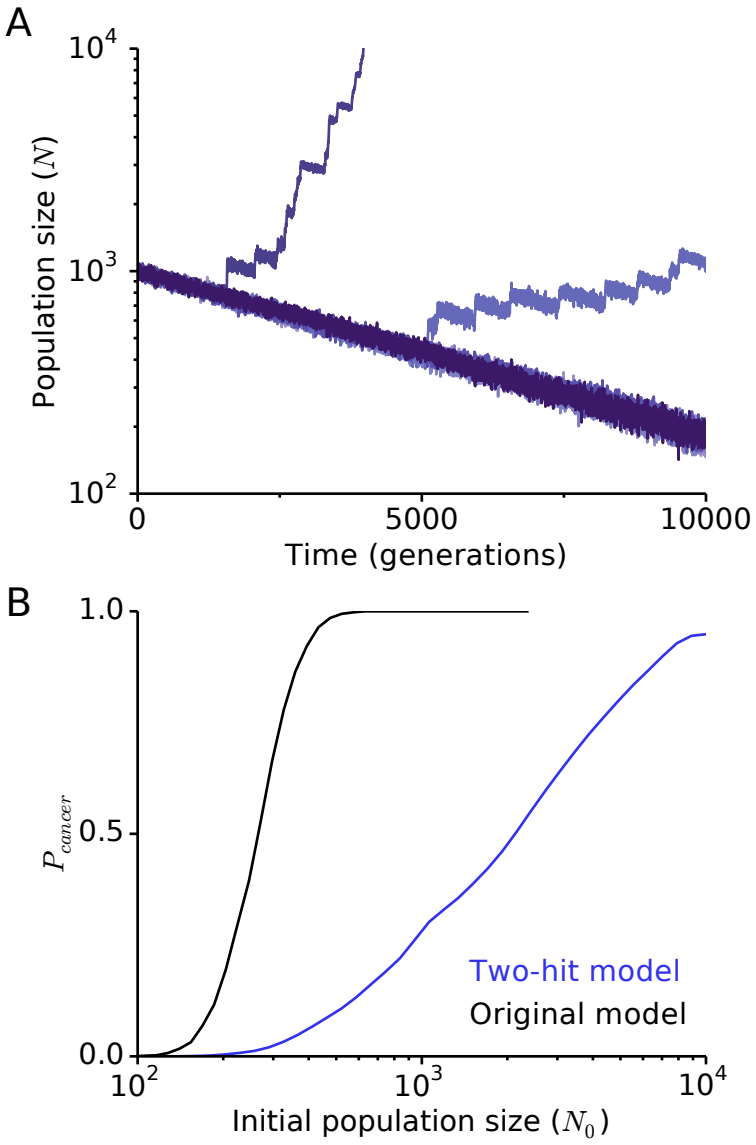


Figure 3.10: (A) We extended our original model to consider cancers where the first driver mutation confers no fitness benefit, the second driver confers a fitness benefit of $2s_d$, and all remaining drivers confer a fitness benefit of s_d. All other properties of the model, including deleterious passengers and a dynamic population size, remain intact. In this figure $s_d = 0.2, s_p = 0.001, \mu_d = 1.4 \times 10^{-5}$, and $\mu_p = 0.1$. Trajectories exhibit a period of stasis where the first driver accumulates (according to neutral dynamics) before the second driver arrives. After this period of stasis, most trajectories progress to cancer. Passengers continue to accumulate and cause extinction in this model. **(B)** Like the original model, we observed a population size-dependent barrier to progression in the two-hit model. However, because of the long and variable period of stasis where passengers continue to accumulate, the critical population size $N^{*,\mathrm{two-hit}}$ is larger and the transition from the extinction and growth regimes is slower. This is more-or-less equivalent to our model with a larger value of s_d and smaller value of T_d.

4

An analytical deconstruction of drivers and passengers

4.1 An analytical model of dynamics

In chapter 3 on page 22, we demonstrate that dynamics are described by two counteracting forces: an upward velocity v_d resulting from accumulating beneficial drivers,

and a downward velocity v_p resulting from accumulating deleterious passengers. The upward velocity v_d was further subdivided into a product of the rate at which new drivers fixate in the population f times their effect on population size once fixated ΔN (figure 3.8 on page 35) [*]. The velocities v_d and v_p are balanced at a critical population size N^*, at which the population is approximately equally likely to go extinct or progress to cancer.

While we were able to describe the average behavior of our population (chapter 3 on page 22), our system, like cancer, is inherently stochastic. Its complete dynamics is best described by a differential equation with stochastic jumps:

$$\begin{aligned} dN &= v_p dt + \Delta N dn_d \\ n_d &=\xrightarrow{f} n_d + 1 \end{aligned} \tag{4.1}$$

In this equation, the change in population size dN is the product of a deterministic component $-v_p$, along with a stochastic component describing the random arrival of new drivers ($\Delta N dn_d$). Below, we use this equation to estimate the probability of cancer for any population size $P_{\text{cancer}}(x)$ and the mean waiting time to cancer $\overline{t_{\text{cancer}}}(x)$, where $x = N/N^0$ is a dimensionless population size briefly discussed in the previous chapter and more thoroughly discussed below. Lastly, we noticed that simulations differed from the formalism we presented in the previous chapter, if the mutation rate μ increases beyond 10^{-8}, or if we explore a broader range of passenger deleteriousness s_p (figure 4.1 on page 46). These discrepancies are resolvable by considering two phenomena neglected in our first derivation: selection against passengers, and passen-

[*] While we assume that drivers arise at random time intervals, this assumption is not always true. Because unfixed passengers can interfere with the fixation of drivers, a driver is more likely to fixate immediately following a previous driver fixation event[57]. Ignoring this caveat does not significantly alter dynamics in the parameter space explored here.

ger's effect on both the fixation probability and clone fitness of drivers. Fortuitously, accounting for these phenomena does not alter equation (4.1) on the preceding page, nor the overall framework of our analytical model. Instead, they only affect the rates v_p, f, and jump size ΔN in our model. Thus, with the refined formalism, we can describe dynamics across a very broad range of parameters (figure 4.1 on page 46). More importantly, we observe drastic reductions in the probability of adaptation at high mutation rates and encompassing moderately deleterious passengers. These findings suggest novel strategies to cancer therapy discussed in chapter 6 on page 88.

From equation (4.1) on the previous page and figure 4.2 on page 47, it is evident that population size is the state variable of our system and, as such, is all that is needed to describe future dynamics. By converting population size into a dimensionless parameter $x = N/N^*$ (and $x^0 = N^0/N^*$), the probability of cancer collapse onto a simple curve $P_{\text{cancer}}(x)$ (figure 3.8 on page 35)—further underscoring the importance of the critical population size. Hence, we will use this dimensionless quantity x heavily throughout the remainder of our analysis.

4.1.1 Estimating the probability of cancer

Using equation (4.1) on page 42 we can describe how the probability of extinction changes in an infinitesimal time due to either passenger accumulation or a rare driver jump:

$$P_{\text{cancer}}(x) = f(x)dtP_{\text{cancer}}[x + \Delta N(x)] + [1 - f(x)dt]P_{\text{cancer}}[x - v_p(x)dt]$$

In this equation, we see that is the probability of cancer at x is the probability of a jump times the probability of cancer after the jump ($f(x)dtP_{\text{cancer}}[x + \Delta N(x)]$) plus the probability of decline times the probability of cancer after the decline ($[1 - f(x)dt]P_{\text{cancer}}[x - v_p(x)dt]$). Note that f, ΔN, and v_p are all functions of x in the equation above. Defining these functions in such a general form makes solving the stochastic differential equation impossible. So we note that each function is approximately linear in x. Thus, we can replace each function with a constant times x: $f(x) \to fx$, $\Delta N(x) \to \Delta Nx$, and $v_p(x) \to v_p x$.

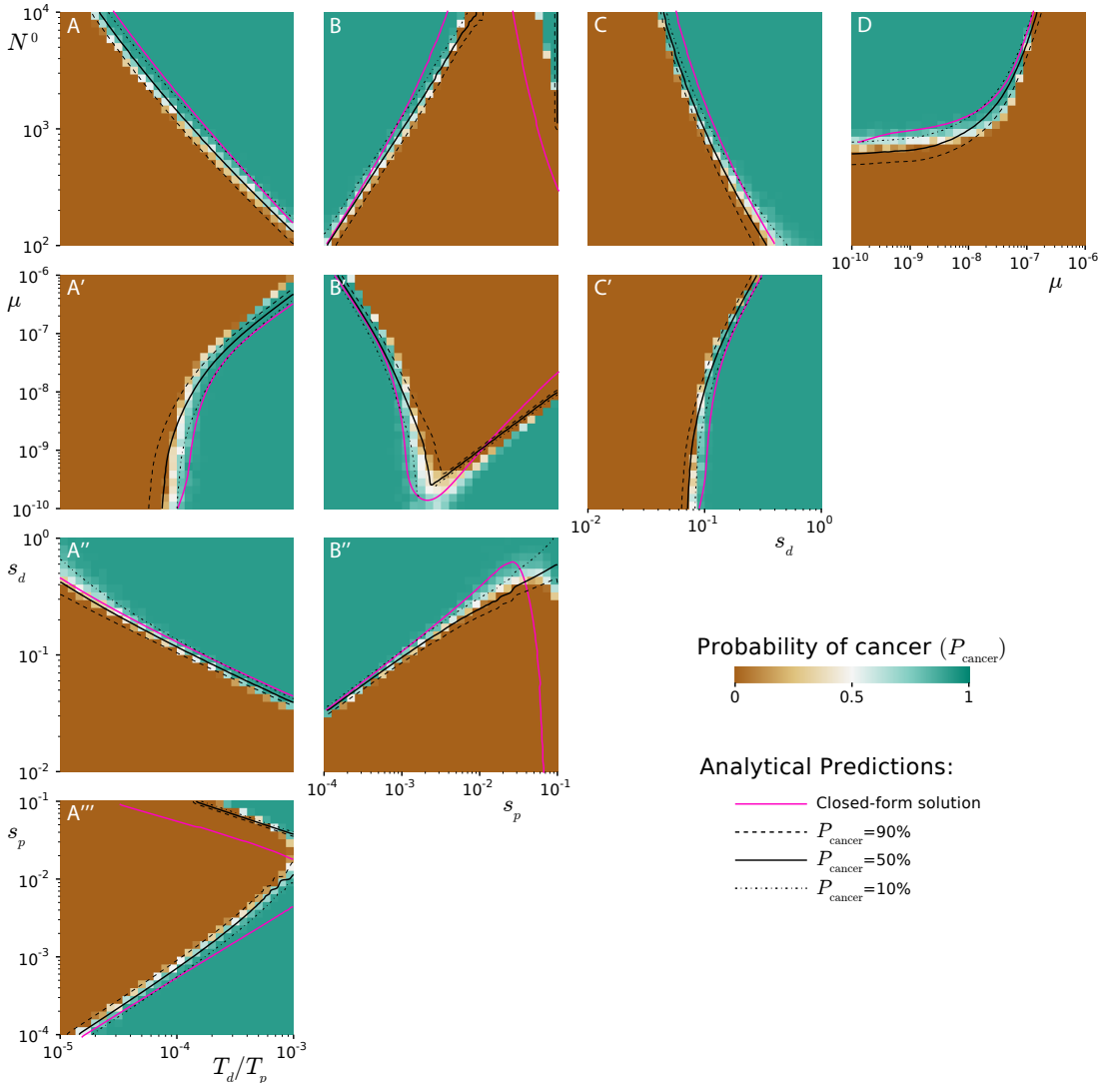
The probability of cancer after a decline can be expanded via a Taylor series: $P_{\text{cancer}}[x - v_p xdt] \approx P_{\text{cancer}}(x) - v_p xdtP'_{\text{cancer}}(x)$. Along with the linear approximations we made for f and θ, this reduces the above equation to:

$$\frac{v_p}{f}P'_{\text{cancer}}(x) = P_{\text{cancer}}(\theta x) - P_{\text{cancer}}(x) \tag{4.2}$$

Here, $\theta = 1 + \Delta N \approx 1 + s_d$ denotes the logarithmic change in population size after a driver jump. Next we notice from simulations that $P_{\text{cancer}}(x)$ changers most significantly when x is near 1. Hence, we can logarithmically-transform x and solve

Figure 4.1: Analytical framework predicts probability of cancer across parameter space. The probability of progression, determined from the outcome of 3,000 simulations (for each data-square) propagated until extinction or rapid growth, across the parameter range of our model. In simulations, we observe parameters where progression occurs, fails, and is rare. A sophisticated analytical framework incorporating selection against passengers, hitchhiking of passengers onto driver mutations, and stochasticity in population size predicts observations well (black lines). This sophisticated analytical model uses two solutions for Muller's Ratchet in various parameter regimes (see section 4.1.3 on page 53), an estimate of the quantity of hitchhiking passengers (see section 4.1.4 on page 57), and a stochastic differential equation of the population size to estimate probabilities of progression (see section 4.1.1 on page 44). A simplified framework, which offers a closed-form solution, is possible and works reasonably well (magenta). This solution differs from the more precise solution in two ways: (1) a novel, simplified estimate of Muller's ratchet is used (Eq. 4.8), and (2) we neglect hitchhikers that accumulate after a new driver clone arises (i.e. $\delta p_{\mathrm{S}} = 0$). **(A-A‴)** P_{cancer} increases for all parameters, as the relative target size of drivers T_d verses passengers T_p increases. **(B-B‴)** and **(A‴)** P_{cancer} exhibits a local minimum verses the selection against passengers s_p. When selection against passengers is very weak, passengers are effectively neutral. When selection against passengers is too strong, natural selection prevents passengers from accumulating. Deleterious passengers are most effective at preventing cancer when moderate in effect size. The local minimum suggests that there may be two types of cancers: those existing in an environment or genetic context where passengers are weak, perhaps buffered by an activated UPR; and those that succeed by exacerbating passengers' deleterious effects, perhaps by decreasing their mutation rate (see **B'**). **(C, C′, A″, B″)** The probability of cancer always increases with s_d. This parameter has a profound affect on cancer progression as it increases the benefit of drivers and their probability of fixation. Hence, the boundary between success and failure appears to be almost independent of the other parameters. **(D)** and **(A′)** An increasing mutation rate affects the probability of cancer very little at first; however, once it exceeds a critical value ($\mu^* = s_d/T_p$, section 4.2 on page 63), the probability of cancer drops precipitously.

Figure 4.1: (continued)



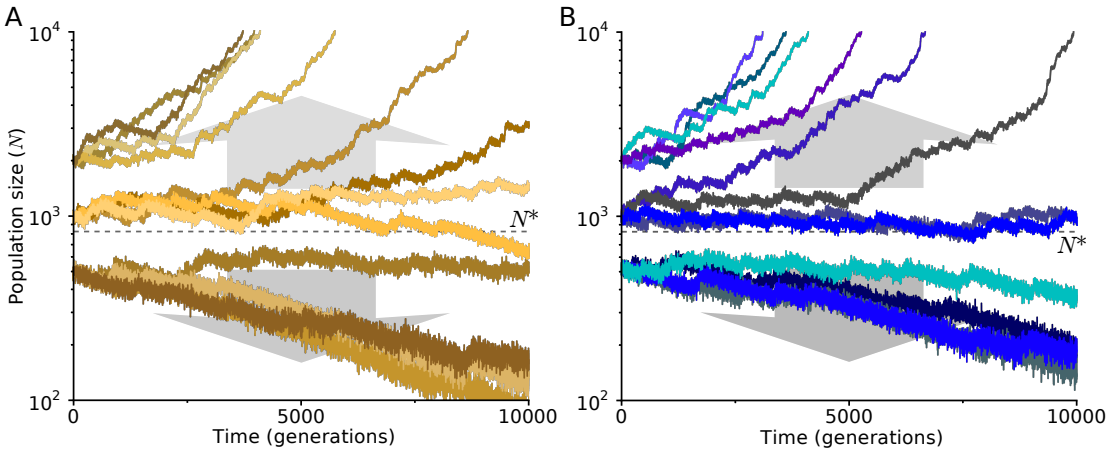


Figure 4.2: Simulations exhibit path independence. **(A)** The 12 trajectories from **Figure 1A**, initiated at $N^0 = \{500, 1000, 2000\}$. **(B)** An additional 12 trajectories, initiated at various N^0, but plotted once they cross $N = \{500, 1000, 2000\}$. Populations that crossed $N = 500$ and $N = 2000$ were initiated at $N^0 = 1000$, while populations that crossed $N = 1000$ were initiated at $N^0 = 500$. Dynamics in **A** and **B** appear identical, demonstrating that populations beginning at different initial sizes N^0 will behave similarly, if they have the same current size. Thus, populations exhibit path independence and can be fully described by one state variable: $x = N/N^0$.

this new variable $y = log(x)$ via a Maclaurin Series:

$$\begin{aligned}\frac{v_p}{f}\frac{dP_{\text{cancer}}(y)}{dy} &= e^y(P(y + \text{Log}(\theta)) - P(y)) \\ &\approx e^y(\text{Log}(\theta)\frac{dP_{\text{cancer}}(y)}{dy} + \frac{1}{2}\text{Log}^2(\theta)\frac{d^2P_{\text{cancer}}(y)}{dy^2} + ...)\end{aligned}$$

Now, by reverting from y back to x, we obtain:

$$\frac{v_p}{f}P'_{\text{cancer}}(x) = \text{Log}(\theta)xP'_{\text{cancer}}(x) + \frac{1}{2}\text{Log}^2(\theta)x^2P''_{\text{cancer}}(x) + \frac{1}{2}\text{Log}^2(\theta)xP'_{\text{cancer}}(x) \quad (4.3)$$

By eliminating the last term in this solution, a reasonable approximation because $\text{Log}^2(\theta) \ll 1$, the differential equation is now solvable. Its boundary conditions (es-

sentially the definitions of *cancer* and *extinction*) are:

$$P_{\text{cancer}}(x = 0) = 0$$

$$P_{\text{cancer}}(x = \infty) = 1$$

The probability of cancer after infinite time is (demonstrable by substitution in equation (4.3) on the previous page):

$$P_{\text{cancer}} = 1 - \gamma(\tfrac{2}{\text{Log}(\theta)}, \tfrac{2}{\text{Log}(\theta)x}) \tag{4.4}$$

Here, $\gamma(s, x) = 1/\Gamma(s) \int_0^x e^{-t} t^{s-1} dt : \Gamma(s) = \int_0^\infty x^{s-1} e^{-x}\ \mathrm{d}x$ is the normalized incomplete gamma function. This solution is parameterized by two dimensionless quantities: θ and x, which represent the jump size in population of driver sweeps and our effective population size respectively.

4.1.2 Estimating the mean time to progression

We can also use equation (4.1) on page 42 to solve for the waiting time to cancer. This can be accomplished in two ways: (1) we can simulate random driver jumps and deterministic passenger decline directly, and (2) we can approximate the mean waiting time to cancer using a Taylor expansion similar to the strategy we employed to solve for the probability of cancer. These two approaches agree with each other (thus, illustrating their accuracy), and offer key insights into the evolutionary parameters that affect age-incidence curves (figure 4.3 on the next page).

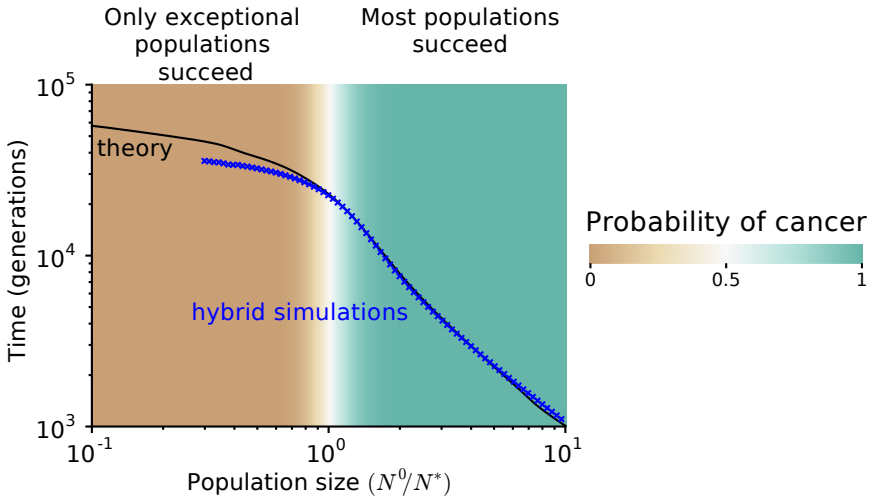


Figure 4.3: Age-incidence curves in our model depend primarily on s_d and match observed age-incidence curves well at mid/late life. **A** Mean waiting time to cancer $\overline{t_{\text{cancer}}}(x)$ decreases as initial population size $x = N^0/N^*$ increases. $\overline{t_{\text{cancer}}}(x)$ was solved from equation (4.1) on page 42 using two methods: (1) stochastic 'hybrid' Gillespie simulations (equation (4.5) on the next page), and (2) an analytical approximation (equation (4.6) on page 52, labeled 'theory'). Agreement between these two estimates suggests that our solution is accurate and a simplified analytical treatment of dynamics is possible (see section 4.1.2 on the previous page for details). These results demonstrate that the shape of our predicted age-incidence cures (below) should depend almost entirely on s_d and not x when $x < 1$, thereby simplifying interpretation of age-incidence curves.

equation (4.1) on page 42 can be simulated using a "hybrid" Gillespie algorithm: a meta-simulation of driver- and passenger-accumulation events that we, originally, observed arising from our atomistic simulations of birth, death, and mutational events. The advantage of this technique is that it allows us to quickly simulate billions of tumors, which would be computationally impossible via full-detail simulations. Because we are confident that we are accurately estimating the rate of driver and passenger accumulation events (figure 4.1 on page 46), this simplification should retain accuracy. To simulate equation (4.1) on page 42 directly, we must consider that the instantaneous probability of a driver jump $f[x(t)]$ is a function of a constantly declining population size due to passenger accumulation: $x(t) = x_{n_d}(1+s_p)^{v_p/s_p t} \approx x_{n_d} e^{-v_p t}$. Here, x_{n_d} is the population size after the last driver jump. Thus, the waiting time between drivers $\Delta t = t_{n_d+1} - t_{n_d}$ is:

$$\begin{aligned} \textstyle\int_0^{\Delta t} f N(t')dt' &= \zeta \\ f \textstyle\int_0^{\Delta t} N_{n_d} e^{-v_p t'} dt' &= \zeta \\ \Delta t &= -\tfrac{1}{v_p}\mathrm{Log}(1 - \tfrac{v_p \zeta}{f N_{n_d}}) \end{aligned} \tag{4.5}$$

ζ is an exponentially-distributed random number with mean 1. Using our precise calculations of f, v_p and ΔN below, we can now simulate equation (4.1) on page 42 directly.

We can also solve equation (4.1) on page 42 for t_{cancer}, using the exact same approximations as we did to estimate $P_{\text{cancer}}(x)$. To do this, we begin with a Master Equation for the probability of acquiring a cancer after waiting time t when currently at size x:

$$P_{\text{cancer}}(x,t) = f(x)\delta t\ P_{\text{cancer}}(\theta x, t+\delta t) + [1 - f(x)\delta t]P_{\text{cancer}}[x - v_p(x)\delta t, t+\delta t]$$

The mean waiting time to cancer is then:

$$\overline{t_{\text{cancer}}(x)} = \int_0^\infty t P_{\text{cancer}}(x,t)dt$$

Before substituting the Master Equation into this definition, we must first utilize a first-order Taylor series expansion about θ and δt:

$$P_{\text{cancer}}(\theta x, t-\delta t) \approx P_{\text{cancer}}(x,t) + \frac{\partial P_{\text{cancer}}(x,t)}{\partial x}(\theta - 1) + \frac{\partial P_{\text{cancer}}(x,t)}{\partial t}\delta t$$

This leads to the solution:

$$\overline{t_{\text{cancer}}}(x) = \int_0^\infty P_{\text{cancer}}(x,t)tdt + \delta t \int_0^\infty \frac{\partial P_{\text{cancer}}(x,t)}{\partial t}tdt + [f\delta t(\theta - 1) + \big(1-f\big)\big(v_p\delta t\big)] \int_0^\infty \frac{\partial P_{\text{cancer}}(x,t)}{\partial x}tdt$$

The first integral in this solution is simply the definition of our mean waiting time ($\overline{t_{\text{cancer}}}(x)$). The second integral can be integrated by parts by noting that $\lim_{t\to\infty} tP_{\text{cancer}}(x,t) = 0$ (otherwise, $\overline{t_{\text{cancer}}}(x)$ would be undefined). Lastly, the third integral reduces to $\overline{t'_{\text{cancer}}}(x)$. Thus, we eventually find:

$$f(x)[\rho_c(\theta x) - \rho_c(x)] - v_p(x)\rho'_c(x) + P_{\text{cancer}}(x) = 0$$

Here, $\rho_c(x) = P_{\text{cancer}}(x)\overline{t_{\text{cancer}}}(x)$. This equation has a nearly identical form to equation (4.2) on page 44. So we used a similar Second-Order Maclaurin series expansion of $\text{Log}(x)$ to approximate its solution:

$$\overline{t_{\text{cancer}}}(x) = \frac{2}{f\text{Log}^2(\theta)}\Big[\int_x^\infty \frac{dy}{y^3}\frac{P_{\text{cancer}}(y)[1-P_{\text{cancer}}(y)]}{P'_{\text{cancer}}(y)} + \frac{1-P_{\text{cancer}}(x)}{P_{\text{cancer}}(x)}\int_0^x \frac{dy}{y^3}\frac{P^2_{\text{cancer}}(y)}{P'_{\text{cancer}}(y)}\Big] \quad (4.6)$$

These integrals can be numerically computed using Simpson's Method and yield a solution that is in good agreement with the hybrid simulations described above (figure 4.3 on page 49).

Our solution for the waiting time to cancer is most illustrative when $x \ll 1$—the regime that we expect to contain most tumors. In this regime, the mean time of cancer progression increases as $-\text{Log}(x)/v_p$, which implies two interesting properties of $\overline{t_{\text{cancer}}}$. First, x has a very weak, sub-linear, effect on the waiting time and does not significantly alter the shape of incidence curves (discussed in chapter 5 on page 68). Second, the waiting time to cancer is dictated by v_p (the accumulation rate of passengers), thus offering yet another reason to continue investigating the rate of deleterious passenger accumulation.

The mean time to cancer $\overline{t_{\text{cancer}}}(x)$ is a quantity that is conditioned on a population actually progressing to cancer. Hence, it depends heavily on the probability of adaptation $P_{\text{cancer}}(x)$. Because $P_{\text{cancer}}(x)$ has an inflection point at $x = 1$, the mean waiting time to cancer $\overline{t_{\text{cancer}}}(x)$ behaves very differently when $x > 1$, than when $x < 1$. When $x > 1$, the mean waiting time is approximately $\overline{t_{\text{cancer}}} \approx \int \frac{1}{<dN/dt>} dN$, i.e. the solution behaves as we would expect from our mean solution—as nearly all

cancers succeed. However, when $x < 1$, the mean waiting time to cancer defies mean behavior, as the mean cancer goes extinct. Only the rare, exceptional populations that progress to cancer are weighted in the mean of $t_{\text{cancer}}(x)$; these exceptional populations grow much faster than the average population. Hence, the increase in waiting time to cancer is sub-linear with respect to x when $x < 1$.

4.1.3 Accumulation of deleterious passengers

Passenger mutations accumulate and drag populations down with a rate v_p. This quantity is a product of passenger's arrival rate $\mu_p N$, their fixation probability π_p, and their effect on population size once fixated Ns_p (i.e. $v_p \approx \mu_p s_p N$, when $\pi_p \approx \frac{1}{N}$). In the previous chapter, we assume that the fixation probability is approximately neutral ($\pi_p \approx 1/N$); however, when selection is stronger that genetic drift, the fixation probability becomes less than the neutral rate. A number of studies have focused on the fixation probability of deleterious mutations in a population, termed Muller's Ratchet[92,48,16,47]. In general, estimates of Muller's Ratchet (and consequentially π_p) begin by considering the distribution of deleterious alleles in a population of infinite size in mutation-selection balance—where allele frequencies are not changing. At equilibrium, such a population exhibits a Poisson distribution in the number of segregating passengers δp within a cell $N_{\delta p}$, defined by a characteristic parameter $\lambda_p = \mu_p / s_p$ (figure 4.4 on the following page):

$$N_{\delta p} = N \frac{e^{-\lambda_p} \lambda_p^{\delta p}}{\delta p!} \tag{4.7}$$

If we then consider a population of finite size, we find that the allele frequencies fluctuate due to genetic drift. If fluctuations in the fittest class ($N_{\delta p=0} = N e^{-\lambda_p}$)

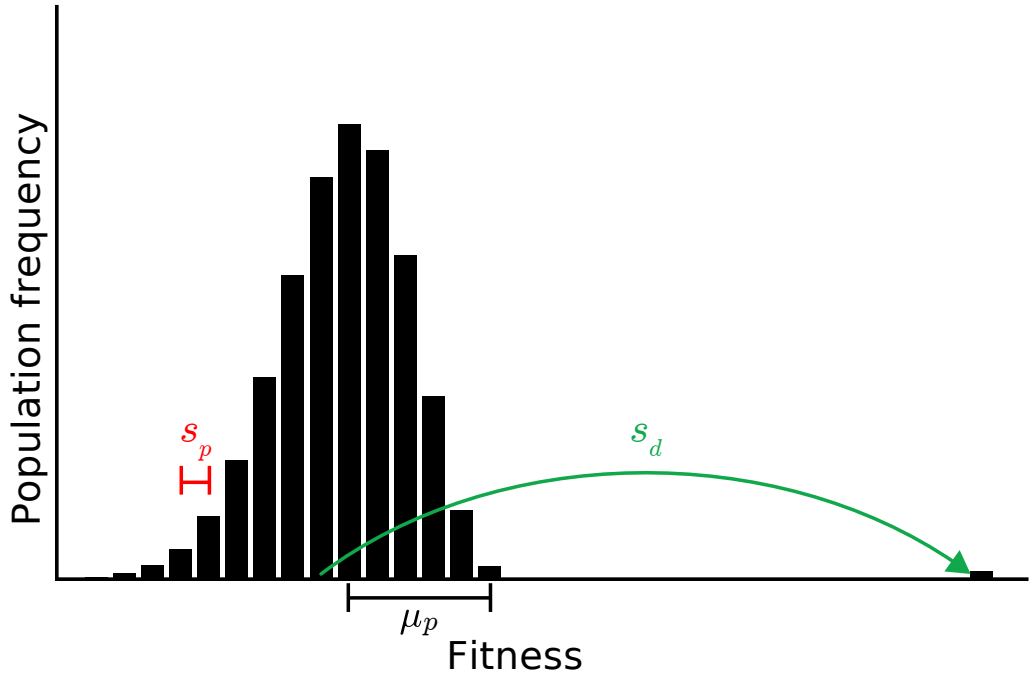


Figure 4.4: Passengers in mutation selection balance. Diagram illustrating how the load of segregating (unfixed) passengers influences relative cell fitness and the probability of a driver fixating. Hitchhiking passengers reduce a driver's fitness benefit and probability of fixation. For a population of infinite size that has evolved to equilibrium, the frequency of individuals with a certain number of segregating passengers abides by a Poisson distribution. The mean reduction in fitness is simply μ_p.

are large enough to cause this fittest class to go extinct, then it is irrevocably lost from the population. This irrevocable loss is considered a 'click' of Muller's Ratchet. The new fittest class—individuals harboring one segregating passenger prior to the 'click'—then relaxes to a new equilibrium that fluctuates, and the process repeats. Estimating the time required for a new fittest class to relax to equilibrium size immediately following a 'click' is non-trivial and dependent upon the parameters of the system: N, s_p, and μ_p, which can vary by orders of magnitude depending upon the evolutionary system in question; hence there are many estimates of the exact rate of Muller's Ratchet.

We present and utilize 3 estimates of the rate of Muller's Ratchet: (1) a solution that works well for most values of s_p, μ_p, and N considered here (figure 4.1 on page 46) and simply ignores the time to equilibration after 'clicks'; (2) a traveling-wave solution accurate for large values of λ_p[16] that allows the distribution of segregating passengers to be far from equilibrium, but presumes that that the size of neighboring mutational classes are uncorrelated; and (3) a solution accurate for small values of λ_p[47] that considers correlations between neighboring fitness classes, but requires that the population be in quasi-equilibrium (i.e. near mutation-selection balance). Estimates (2) and (3) accurately describe Muller's Ratchet across complimentary regions of our phase space. By combining these later two estimates with estimates of the number of hitchhiking passengers and their effects on the probability of driver fixation events, we developed a precise description of our model's dynamics (figure 4.1 on page 46 black lines).

If we simply ignore the equilibration time of a population into mutation-selection balance, then we can estimate the rate of Muller's Ratchet with a closed form solution that is applicable to all values of s_p, μ_p, N investigated here. We assume that the

probability of a 'click' is approximately the probability of a new passenger fixating *within* the fittest class: $N_{\delta p=0} = Ne^{-\lambda_p}$. In other words, to a first-approximation, deleterious passengers simply reduce the effective population size of our system, such that $N_e \sim Ne^{-\lambda_p}$. The probability of a lone deleterious allele fixating within this fittest class is describe by a Moran Process[95]. Hence,

$$\pi_p^{(1)} = \frac{s_p}{(1+s_p)^{N_e} - 1} \tag{4.8}$$

This refined fixation probability $\pi_p^{(1)}$ is then used to correct the downward velocity due to passengers, using the same formula for v_p derived in the main text:

$$v_p^{(i)} = \mu_p s_p N \pi_p^{(i)} \tag{4.9}$$

This equation links v_p to the passenger fixation probabilities calculated above, and the other two fixation probabilities calculated below.

The solution for Muller's Ratchet as a traveling wave, which we apply when $\lambda_p < 1$, was obtained from[16]:

$$\frac{\mathrm{Log}(\frac{Ns_p}{\sqrt{\lambda_p}})}{\lambda_p} \approx 1 - \frac{\pi_p^{(2)}}{2}[\mathrm{Log}^2(\frac{e}{\pi_p^{(2)}}) + 1] - \frac{1}{\lambda_p}\mathrm{Log}[\frac{(\pi_p^{(2)})^{3/2}}{\sqrt{1 - pi_p^{(2)}}} \frac{\mathrm{Log}(\frac{e}{\lambda_p})}{1 - \pi_p^{(2)}\mathrm{Log}(\frac{e}{\lambda_p}) + \frac{5}{6\lambda_p}}] \tag{4.10}$$

Because this equation is transcendental, we solved for $\pi_p^{(2)}$ using Brent's Method.

When $\lambda_p \geq 1$, a quasi-stationary analysis of the mutation classes becomes appropri-

ate. This analysis was first done in ref.[47], resulting in a solution of the form:

$$T_{click} = \frac{e-1}{s_p} e^{\frac{s_p N_{\delta p=0}}{2(e-1)}} \tag{4.11}$$

The fixation probability is then simply the inverse of the 'click' time: $\pi_p^{(3)} = 1/T_{click}$.

Lastly, there is a discontinuity between the above two solutions at their intersection: $\lambda_p = 1$. We resolved this by interpolating between the two solutions, as follows:

$$\pi_p^{(\text{combined})} = \lambda_p \pi_p^{(2)} + (1-\lambda_p)\pi^{(3)}$$

In essence, we have utilized three different solutions for the accumulation of deleterious passengers, applicable in various parameter regimes: a neutral approximation utilized in chapter 3 on page 22, a traveling-wave solution, and an estimate based on fluctuations in Muller's Ratchet. Figure 4.5 on the following page illustrates the accuracy of these three solutions and their applicability by highlighting their domain of intended use and comparing them to simulations without advantageous drivers.

4.1.4 Effects of deleterious passengers on fixation probability and clone fitness of drivers

The occurrence and fixation of driver mutations are rare events, separated by nearly random time intervals, with a frequency of occurrence $f = \mu_d N \pi_d$. Here, π_d is the fixation probability of a new mutant driver once it arises in the population. In the mean velocity model presented in the previous chapter, we estimate that $\pi_d = s_d/(1+s_d) \approx s_d$. However, this result assumes that there are no other non-neutral alleles in the population. In reality, there are many segregating passengers in the population, and potentially other segregating drivers.

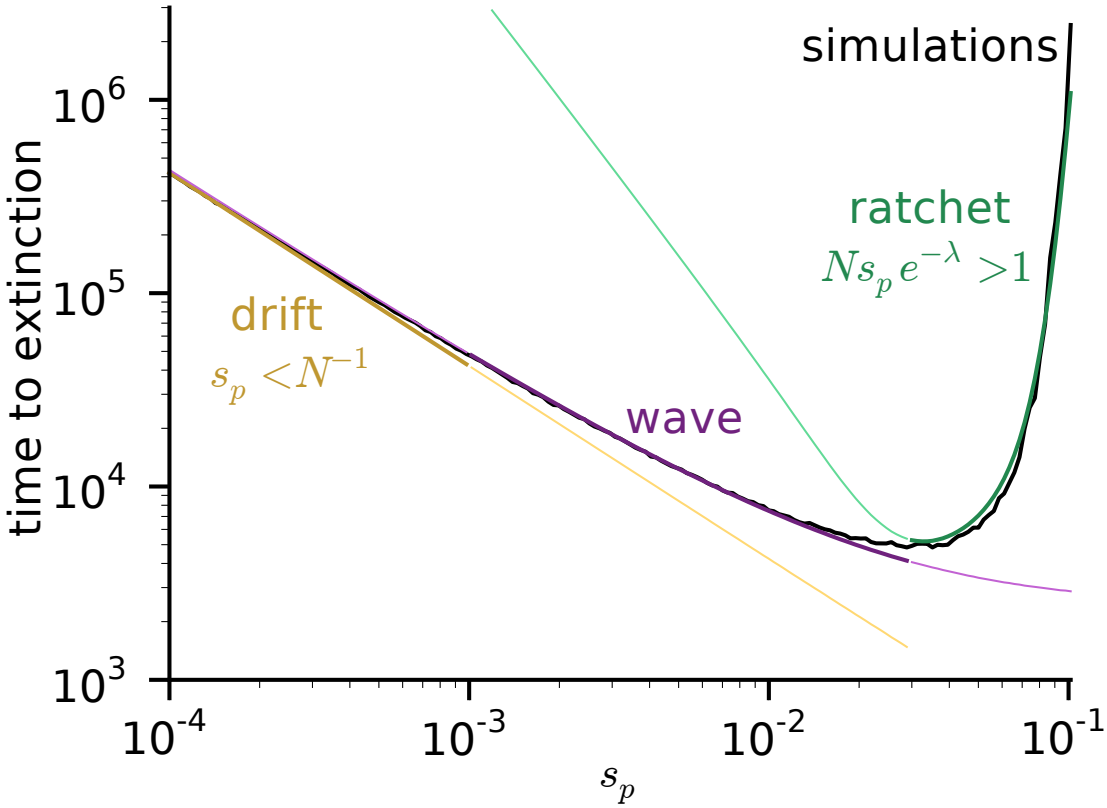


Figure 4.5: Analytical estimates of passenger accumulation. To test the accuracy of our three analytical approximations of passenger accumulation: Muller's Ratchet, a traveling wave, and a neutral model (drift); we simulated cancer populations with estimated parameters, except driver mutations were impossible. All simulations progressed to extinction via passenger accumulation, but did so with varying time. By assuming $N(n_p) = N^0(1 + s_p)^{-n_p}$ and that a population goes extinct once $N = 15$ (i.e. stochastic fluctuations kill very small populations), we could analytically estimate the time to extinction. Our analytical approximations agree with the simulations for their regimes of validity (heavy lines). For completeness, thinned and lighted predictions of the analytical results outside of their intended domain are presented.

The presence of other drivers in the population, which interfere with the fixation of our clone of interest, is a phenomena commonly described as *Clonal Interference*[46]. Clonal Interference becomes significant in the population once the time required for a driver to fixate [$\sim \text{Log}(N)/s_d$ generations] approaches the fixation rate ($f \approx \mu_d N s_d$). Nascent precancerous population are in a space of evolutionary parameters where Clonal Interference is particularly negligible: population size is small ($N \sim 10^3$), and drivers are rare ($\mu_d \sim 10^{-5}$), but strong ($s_d \sim 10^{-1}$). Thus, we do not consider its effects here. However, for a larger tumor population, clonal interference may become very significant. This is especially true in a poorly-mixed population, where beneficial alleles take longer to sweep through the population[66].

Segregating passenger mutations can also interfere with a driver sweep by 'hitchhiking' on the expanding clone[57,5]. Hitchhiking passengers are critical for understanding population dynamics at high mutation rates, so we will now consider their effects in detail. Most of the analysis we present has already been performed in two previous works[57,5], however it is integral to our derivation of the critical mutation rate and must be extend to very high mutation rates, so we have decided to repeat much of their work here.

For mathematical analysis, we disentangle two types of hitchhikers: (1) those that reside in the Initial clone before the new driver arises (denoted δp_{I}), and (2) those that arise and fixate in the new driver clone as it Sweeps through the population (denoted δp_{S}). It is necessary to distinguish hitchhikers this way because only the initial hitchhikers (δp_I) significantly alter the fixation probability f, while both types alter the effect size ΔN. The hitchhikers that accumulate during the sweep will generally arise after the clone is of appreciable size; however, once the driver clone is of appreciable size, it is exceedingly likely that it will fixate so long as it remains the fittest

clone in the population.

Here, we consider only the average number of hitchhikers in a driver sweep ($\overline{\delta p_{\mathrm{I}}}$ and $\overline{\delta p_{\mathrm{S}}}$), rather than their entire distribution of quantities; estimates of the average number of hitchhikers appear to explain dynamics reasonably well (first shown in[57] and also evident from our analysis' good agreement with simulations in figure 4.1 on page 46). Thus the probability that a new clone fixates in the absence of Clonal Interference is (figure 4.4 on page 54):

$$\pi_d(\overline{\delta p_{\mathrm{I}}}) = \frac{s'_d(\overline{\delta p_{\mathrm{I}}})}{1 + s'_d(\overline{\delta p_{\mathrm{I}}})} : s'_d(\overline{\delta p_{\mathrm{I}}}) = s_d - \overline{\delta p_{\mathrm{I}}} s_p \tag{4.12}$$

The jump size ΔN becomes:

$$\Delta N' = N[s_d - (\overline{\delta p_{\mathrm{I}}} + \overline{\delta p_{\mathrm{S}}}) s_p] \tag{4.13}$$

We can conclude our analysis of hitchhikers once we obtain $\overline{\delta p_{\mathrm{I}}}$ and $\overline{\delta p_{\mathrm{S}}}$. These quantities were first derived in[57]. We use their results (summarized below), along with a minor necessary adjustment for populations when λ_p is large, to complete our analytical model of cancer progression.

For a new driver clone to take over the population and fixate, it has been shown that its fitness must be greater than the fittest class in the population[57]. This imposes a maximum on the number of initial hitchhikers $\delta p_{\mathrm{I}}^{\max}$ that a successful driver clone can have:

$$\begin{aligned} s_d &> \delta p_{\mathrm{I}} s_p \\ \delta p_{\mathrm{I}}^{\max} &= \lfloor s_d / s_p \rfloor \end{aligned}$$

A clone that does not satisfy this constraint may proliferate for a while in the pop-

ulation, but it will nevertheless be eventually out-competed by fitter clones. When the mean number of hitchhiking passengers (λ_p) approaches this maximum, hitchhikers dramatically reduce both f and ΔN, thus increasing N^* to untenable sizes. This occurs when:

$$\begin{aligned} \lambda_p &= \delta p_{\mathrm{I}}^{\max} \\ \mu_p/s_p &= \lfloor s_d/s_p \rfloor \\ \mu_p &\approx s_d \end{aligned} \tag{4.14}$$

Hence, our analysis suggests a limit on the maximum mutation rate that an adapting population can tolerate: $\mu_p^* \approx s_d$. In simulations, we observe extinction slightly above this threshold (figure 4.6 on page 64). This mechanism of collapse, where populations go extinct by failing to acquire new advantageous mutations or adaptations, differs from the traditional model of mutational meltdown. In the traditional model, advantageous mutations are generally ignored and meltdown occurs only because deleterious mutations accumulate too quickly. In our model, however, traditional mutational meltdown is difficult because populations also acquire advantageous mutations faster as the mutation rate increases. Moreover, traditional meltdown occurs only when the population size is small, making it impossible to occur in a large population like cancer. Our discovery of a new mechanism of meltdown that is independent of population size suggests that mutational meltdown may be induced via cancer therapeutics.

The number of initial segregating passengers in a clone when a driver arises (δp_I) can be obtained by considering, once again, the population at mutation selection balance, i.e. equation (4.7) on page 53. The average number of initial hitchhiking passengers is simply the average of the likelihood of a driver arising in each mutational

class, conditional on the driver successfully sweeping through the population:

$$\begin{aligned} P(\delta p_{\mathrm{I}}) &= \tfrac{1}{\mathscr{N}} N_{n_p=\delta p_{\mathrm{I}}} \pi_d(\delta p_{\mathrm{I}}) \\ \overline{\delta p_{\mathrm{I}}} &= \tfrac{1}{\mathscr{N}} \textstyle\sum_{\delta p_{\mathrm{I}}=0}^{\delta p_{\mathrm{I}}^{\max}} P(\delta p_{\mathrm{I}}) \pi_d(\delta p_{\mathrm{I}}) \\ &= \tfrac{1}{\mathscr{N}} \textstyle\sum_{\delta p_{\mathrm{I}}=0}^{\delta p_{\mathrm{I}}^{\max}} \frac{e^{-\lambda_p} \lambda_p^{\delta p_{\mathrm{I}}}}{\delta p_{\mathrm{I}}!} \frac{s'_d(\delta p_{\mathrm{I}})}{1+s'_d(\delta p_{\mathrm{I}})} \end{aligned} \tag{4.15}$$

Here, $\mathscr{N} = \sum_{\delta p_{\mathrm{I}}=0}^{\delta p_{\mathrm{I}}^{\max}} \pi'_d(\delta p_{\mathrm{I}})$ is a normalization constant.

The above solution fails when λ_p is large. In this circumstance, the population is far from mutation-selection balance. Rectifying the solution in this case is difficult to do precisely, however a simple correction to equation (4.15) can crudely ameliorate the estimate. Because the assumption of mutation-selection balance fails only once the expected number of passengers in the fittest class becomes very small ($N_{n_p=0} = Ne^{-\lambda_p} \sim 1$), we propose that the actual fittest surviving class in the population is the first class of passengers with an expected population size that is greater than the size of fluctuations in the population. Because the variance in a birth and death process is the sum of the rates ($2N$ in our model), the Fittest Surviving Class k_{FSC} is:

$$\begin{aligned} k_{\mathrm{FSC}} &= min_{n_p}[N_{n_p} > \sqrt{2N}] \\ k_{\mathrm{FSC}} &= min_{n_p}[e^{-\lambda_p} \lambda_p^{n_p}/n_p! > \sqrt{\tfrac{2}{N}}] \end{aligned}$$

The corrected distribution of δp_{I} then becomes:

$$\overline{\delta p_{\mathrm{I}}} = \frac{1}{\mathscr{N}} \sum_{\delta p_{\mathrm{I}}=0}^{\delta p_{\mathrm{I}}^{\max}} P(k = \delta p_{\mathrm{I}} + k_{\mathrm{FSC}} | \lambda_p) \pi_d(\delta p_{\mathrm{I}})$$

This simple correction yields a final solution for P_{cancer} that agrees with simulations well (figure 4.1 on page 46).

Lastly, the number of passengers that accumulate during the selective sweep (δp_{S})

can be calculated using a recursive relationship. This relationship begins with the probability of accumulating the maximum possible passengers during the sweep $\delta p_{\mathrm{I}}^{\max}$[57]:

$$\begin{aligned} P(\delta p_{\mathrm{S}} = \delta p_{\mathrm{I}}^{\max}) &= \tfrac{1}{\mathscr{N}_2} \delta p_{\mathrm{I}}^{max} \\ P(\delta p_{\mathrm{S}} = k) &= \tfrac{1}{\mathscr{N}_2} \tfrac{k + s_p P(\delta p_{\mathrm{S}} = k+1)}{1+s_p} \\ \overline{\delta p_{\mathrm{S}}} &= \tfrac{1}{\mathscr{N}_2} \sum_{\delta p_{\mathrm{S}}=0}^{\delta p_{\mathrm{I}}^{\max}} P(\delta p_{\mathrm{S}}) \end{aligned}$$

Where $\mathscr{N}_2 = \sum_{\delta p_S=0}^{\delta p^{max}} P(\delta p_S)$ is a second normalization constant.

4.2 A critical mutation rate

By simulating cancer progression over a broad range of evolutionary parameters (figure 4.1 on page 46), we observed another barrier to cancer: a critical mutation rate, above which the probability of cancer is exceedingly low (figure 4.6 on the following page). Through further analytical treatment (section 4.1.4 on page 57), we found that this mutational barrier is created and determined by the load of segregating (unfixed) passengers in the population.

The origin of a critical mutation rate can be understood by considering the number of segregating passengers per cell, previously shown to be Poisson distributed with mean μ_p/s_p, during mutation-selection balance[53]. The average fitness reduction of a cell due to this mutational load is then μ_p. If this fitness reduction exceeds the benefit of a new driver ($\mu_p > s_d$), then drivers seldom fixate[57] (figure 4.4 on page 54). Hence, cancer is extremely rare above the critical mutations rate $\mu^* = s_d/T_p$.

This analysis offers a new mode by which mutational meltdown operates. Whereas prior models of mutational meltdown consider deleterious mutations in isolation[77], we find that mutational meltdown can occur when deleterious mutations inhibit the

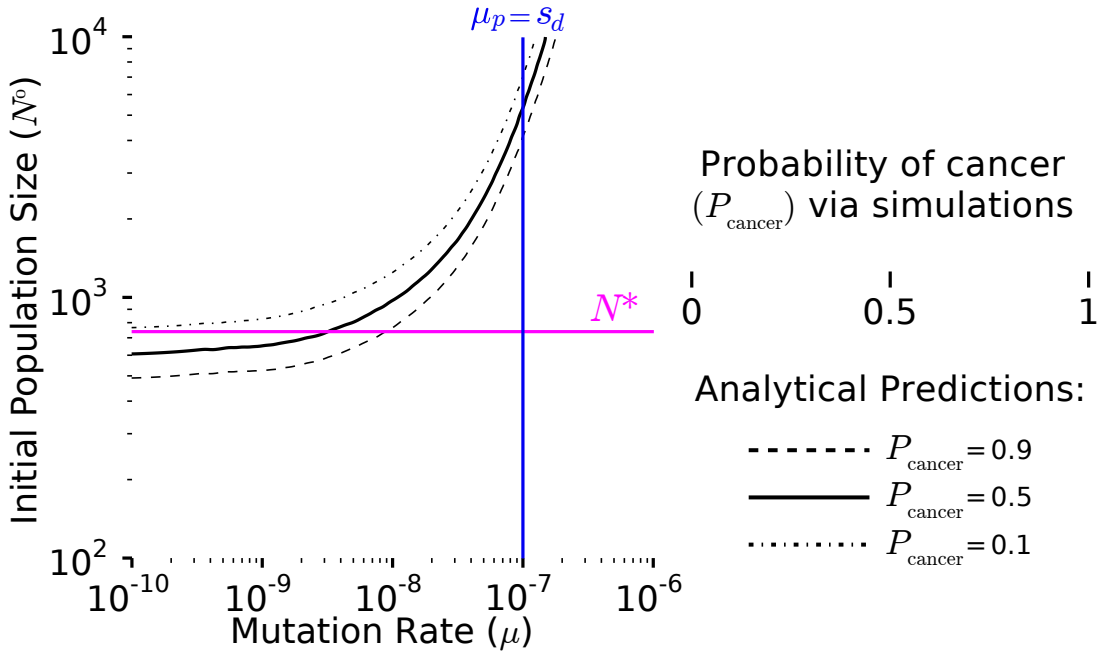


Figure 4.6: Effect of mutation rate on cancer dynamics. The probability of cancer (adaptation) computed by simulations across mutation rates and initial population sizes. Evolutionary parameters partition adaptation into a regime where cancer is almost certain (green), and a regime where it is exceedingly rare (brown), accurately predicted by theoretical estimates of N^* (magenta) and μ^* (blue). A theory incorporating passenger interference with driver sweeps and selection against passengers explains simulations well (black lines).

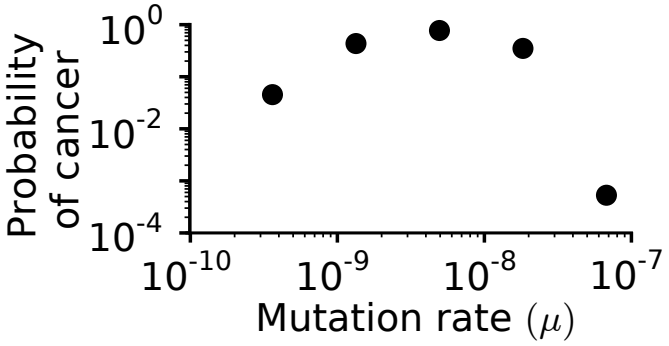


Figure 4.7: An optimal mutation rate for cancer progression. Probability of cancer constrained to grow within a human lifespan ~ 60 years, 10^4 generations, with $N^0 = 10^3$ for various mutation rates. Probabilities exhibit an optimum mutation rate that is near literature estimates of the mutator phenotype in cancer.

accumulation of advantageous mutations. Although previous research found that deleterious mutations can interfere with the fixation of beneficial alleles[57,27], this phenomenon has never been studied in the context of population survival.

When cancer progression is constrained to develop within a human lifetime, we observe an optimum mutation rate for the probability of cancer ($\mu^{\text{opt}} = 10^{-9} - 10^{-8}\text{nucleotide}^{-1} \cdot \text{generation}^{-1}$, figure 4.7), similar to experimentally-measured rates in cancer of 10^{-8} [7]. Above μ^{opt} population meltdown is very common, while below μ^{opt} progression is too slow.

4.3 Estimating the accumulation rate of passengers of varying effect

When deleterious mutations are drawn from a distribution of effect sizes, estimating their accumulation rate becomes more complicated and rigorous analytical treatment of this situation has not been published to our knowledge. However, we will offer an approximate treatment in the limit of small λ here. Figure 3.6 on page 31 demonstrates that the population fitness distributions remain constant between the fixed-

effect and variable effect models, with one caveat: the mean and variance of fitness relative to the fittest class are now $\bar{\lambda} = \mu_p / \bar{s_p}$. Here, $\bar{s_p}$ represents the mean deleteriousness of a passenger. While $\bar{\lambda}$ is relatively large in figure 3.6 on page 31 (approximately 50), this congruence should presumably hold so long as mutation effect sizes are generally smaller than the width of fitness distribution. Obviously, if the variance of the distribution of fitness were very large or undefined (e.g. the variance of a power-law distribution) this approximation may not hold. So long as this is not the case, a new mutation, with its particular fitness effect $\bar{s_p}$, has in the limit of small $\bar{\lambda}$, a fixation probability Π of:

$$\Pi = \frac{\hat{s_p}}{1 - (1 + \hat{s_p})^{N\mathrm{Exp}[\bar{\lambda}]}}$$

Here we used the fixation probability in the weak mutation limit, but with population size described by the size of the fittest class.

For small $\bar{\lambda}$, it is clear that mutations in the deleterious tail, where $\hat{s_p} > \bar{s_p}$, will fixate in this model more often than in a model where all passengers exhibit an effect size of s_p. Indeed, in the fixed-effect model, as mutations become more deleterious they (1) are less likely to fixate in a fittest class of equivalent size, and (2) increase the size of this fittest class—further reducing their probability of fixation. However in a distribution of effect sizes, a particularly-deleterious passenger only alters one half of this equation: the probability of fixating in the fittest class, not the size of the fittest class. The size of the fittest class is approximately defined by the average passenger ($\bar{s_p}$) and not any particularly deleterious passenger $\hat{s_p}$. Thus, the fixation probability of particularly-deleterious passengers declines less rapidly than one might expect from the fixed-effect model (figure 3.6 on page 31).

For large $\bar{\lambda}$, accumulation rates are high and well approximated by a neutral model,

so we expect at most a modest effect of $\hat{s_p}$ relative to $\bar{s_p}$ on the rate of accumulation, although we have and know of no existing theory to accurately describes this scenario. Obviously, the fixation rate should still exhibit some decline, if the fitness distribution has a long deleterious tail—as illustrated in figure 3.6 on page 31.

When considering the effects of treatment strategies on passengers of varying effect, it is important to keep in mind that most of the passengers residing in a cancerous cell have fixated (there are about as many segregating passengers as fixated passengers, but the segregating passengers are often at low frequencies). Therefore, nearly all of the effects of increased selection against passengers will be on mutations that cannot change in frequency in the population. For this reason, cancers grown under previously estimated parameters with passengers drawn from an exponential distribution exhibited similar relapse rate as cancers grown under the fixed effect model.

5

Genomic & Epidemiological evidence for deleterious passengers

In the first two chapters, we outlined a new evolutionary model that included deleterious passengers in a reasonable way, and then found that deleterious passengers can accumulate and dramatically affect progression. In the previous chapter, we investigated the mechanisms by which these passenger accumulated, which also specified

the necessary ingredients for their accumulation and impact on progression. This required formulating a simplified mathematical description of progression. The mathematical framework that emerged now provides us with a clearer understanding of our model's critical features and how it differs from previous models of progression. This now makes comparison to clinical, epidemiological, and genomic observations easier because our solution is robust to evolutionary parameters and contains only two free parameters (N^* and s_d), , exists in closed-form and is easily regressed onto data, is mathematically precise, and easily intuited.

Thus we can now make several, bold, precise, measurable, and distinctive predictions about existing cancers. Considering that we expected each individual passenger to change the growth rate of a cancer cell by about one part in a thousand, this may not have been expected a priori. This accomplishment is a testament to the utility of population genetics theory.

Our model makes several testable predictions:

1. Accumulated passengers in cancer populations can be deleterious to cancer cells;

2. The deleterious effect of an individual passenger has little bearing on its likelihood of accumulation;

3. Drivers that accumulate should have larger effects on phenotype than passengers;

4. A tug-of-war between drivers and passengers in individual cancers;

5. The presence of an effective barrier to cancer that lesions seldom overcome,

6. which results in a broad distribution in the number of accumulated drivers and passengers in tumors.

Recent cancer genomics and epidemiological data provide opportunities to test these predictions. These data systematically supports the above predictions, suggesting that our model is applicable to cancer progression. The data also provides an opportunity to quantitatively measure the fitness effects of both drivers and passengers.

5.1 Passenger mutations observed in cancer exhibit signatures of damaging phenotypes

First, we test whether nonsynonymous passengers found in cancer are damaging or neutral to protein function using comparative genomics. Second, we test whether selection acting against passengers is effective at preventing fixation or largely ineffective, as suggested by our simulations. We analyzed 116,977 cancer mutations curated by the Catalogue of Somatic Mutations in Cancer (COSMIC) and The Cancer Genome Atlas (TCGA). We the classified them as driver and passenger mutation groups and then characterized their effects using PolyPhen, a tool widely used in population and medical genetics to predict the damaging effect of missense mutations[14]. Passengers were identified as missense mutations that show no recurrence and affect genes not listed in a census of possible cancer-causing genes (section A.1 on page 105).

Our method of classifying drivers and passengers throughout this chapter was based on previous techniques employed by the COSMIC and TCGA projects and fully described in the Appendix (see chapter A on page 105). In short, we used the COMIC methodology for pan-cancer analysis, and TCGA methodology for subtype-specific

analysis. In general, their techniques are similar and both utilize the same sources of information: primarily recurrence of mutations across tumor samples, but also both leverage genomic and literature data for further refinement.

The $\Delta PSIC$ metric of PolyPhen measures the degree of evolutionary conservation of a mutated residue[116] by calculating the negative log-likelihood of observing a specific mutation, given the evolutionary history of the protein. Specifically, a mutation with a $\Delta PSIC$ of 1 is e $(= 2.71...)$ times less likely to be observed than the wild-type allele, as computed from a multiple alignment. Thus, a mutation with high $\Delta PSIC$ is more likely to be under natural selection and affecting to molecular function[116], as high values imply that the mutation disrupts a well-conserved residue. PolyPhen has been extensively tested and benchmarked[14].

Figure 5.1 on the following page presents this analysis for passengers, drivers, and three reference datasets: (i) common human missense SNPs; (ii) simulated de novo mutations (randomly generated using a cancer-specific three-parameter model; see section A.1.1 on page 107); and (iii) damaging, pathogenic missense mutations causing human Mendelian diseases (from the Human Gene Mutation Database). As expected, common SNPs are benign and exhibit small $\Delta PSIC$ values, whereas disease-causing mutations, with known damaging effect, exhibit large $\Delta PSIC$ values. Driver mutations exhibit similarly high values of $\Delta PSIC$, significantly greater than randomly generated mutations, indicating that drivers tend to occur at well-conserved loci. From a biochemical perspective, this result shows that, to activate an oncogene or to disable a tumor suppressor, the driver mutation must change a critical and well-conserved residue, e.g. the GTP binding site of Ras or DNA binding domain of p53. From an evolutionary perspective, the conservation of residues that promote tumorigenesis when mutated suggests strong natural selection against the early de-

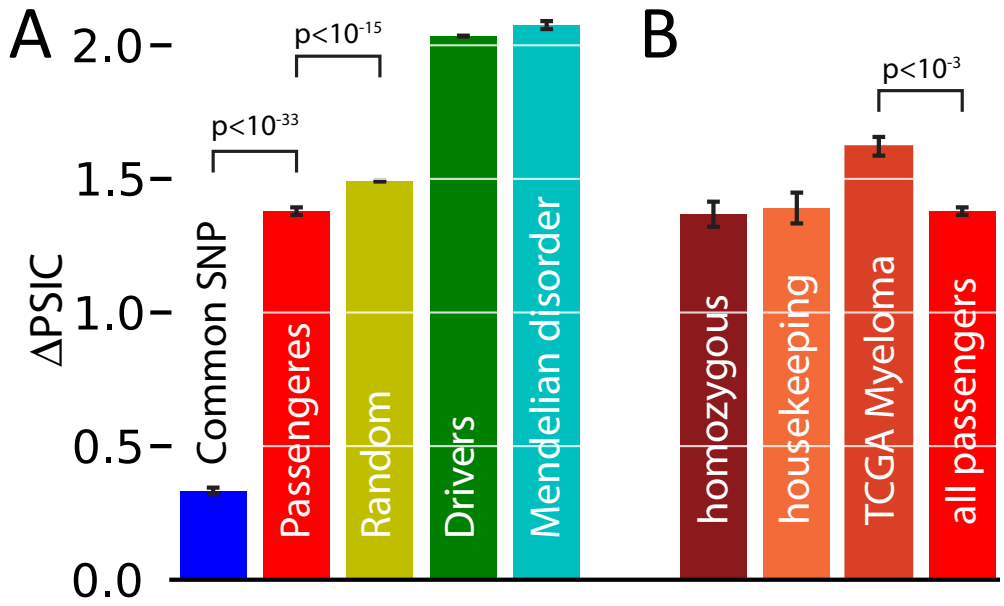


Figure 5.1: Characterization of missense mutations in cancer sequencing data. **(A)** Mutations were assayed using the $\Delta PSIC$ score of PolyPhen, which estimates the damaging effect of a new mutation, given known homologs; mutations with high $\Delta PSIC$ scores are most likely damaging [116]. Passengers have large $\Delta PSIC$, close to random mutations, suggesting that they are deleterious. Drivers have very large $\Delta PSIC$, indicative of their highly non-neutral phenotype. **(B)** Deleterious passenger phenotypes were observed in all subsets of passengers studied, arguing that these results cannot be explained by recessive phenotypes, or lack of expression, or database biases.

velopment of cancer. The ability of $\Delta PSIC$ score to identify drivers as having highly non-neutral phenotypes (i.e., damaging or altering molecular function) also validates its use for characterizing somatic cancer mutations. The exceptionally high $\Delta PSIC$ scores for these mutations are consistent with our third prediction that drivers must be of strong effect.

Most importantly, passenger mutations exhibit $\Delta PSIC$ values that are on average much greater than neutral mutations (figure 5.1; $p < 10^{33}$); therefore, many passengers affect conserved residues and are likely damaging to protein function. This result clearly demonstrates that passenger mutations are non-neutral. To ensure that our set of putative passenger mutations was not contaminated by drivers, we increased

our stringency of passenger classification, but found no statistically significant change ($p = 0.69$) in mean $\Delta PSIC^*$. Additional safeguards are explored below.

Passengers tended to exhibit $\Delta PSIC$ values that were much lower than drivers, supporting the assumption of our evolutionary model that deleterious passengers are generally much weaker than drivers ($s_p < s_d$). The $\Delta PSIC$ values of passengers are close to, but lower than, values of randomly generated mutations (figure 5.1 on the preceding page; $p < 10^{-15}$), suggesting that many passenger mutations evade purifying selection. Still, a statistically-significant difference between these two sets demonstrates slight negative selection against the most deleterious passengers. This comparison of passengers and random mutations fully supports our model's prediction that selection against moderately deleterious passengers is largely ineffective in neoplastic progression (figure 3.2 on page 24).

To rule out possible caveats where passengers have damaging effects on protein function but no effect on the fitness of cancer cells, we performed additional tests in figure 5.1 on the preceding page. For example, passengers with deleterious scores could affect only genes that are functionally unimportant or not expressed in cancer cells. Thus, we considered only passengers in essential and ubiquitously expressed housekeeping genes, but still observe equally high $\Delta PSIC$ scores. This eliminates the possibility that damaging passengers are not expressed or present in unimportant genes. Alternatively, perhaps only recessive heterozygous passengers exhibit high $\Delta PSIC$ scores; if so, cell fitness would remain unchanged because the other allele provides sufficient functionality. We observe equally high $\Delta PSIC$ scores for homozy-

*This stringent classification required that all mutations be confirmed by additional Sanger sequencing and required that all passenger genes not only be non-drivers, but also exhibit a ratio of non-synonymous to synonymous mutations consistent with purifying selection (section 5.1.1 on the next page).

gous passengers (which can arise via LOH events or chromosomal losses), refuting this possibility. Collectively, our analyses show that signatures of damaging mutations are ubiquitous in known passengers and likely affect the fitness of cancerous cells.

5.1.1 Passengers are very rarely eliminated by natural selection

We assayed for signatures of selection in driver and passenger genes by comparing the observed ratio ω of nonsynonymous to synonymous mutations to the predicted ratio using a random model of mutations (section A.1.1 on page 107, figure 5.2 on the following page). Operationally, we defined the ratio of nonsynonymous to synonymous mutations as follows:

$$\begin{aligned}\omega &= \frac{\omega_{\text{observed}}}{\omega_{\text{expected}}} \\ &= \frac{O_{\text{nonsynonymous}}/O_{\text{synonymous}}}{E_{\text{nonsynonymous}}/E_{\text{synonymous}}}\end{aligned}$$

Here O represents the observed mutations within a gene of interest (synonymous or nonsynonymous) and E represents the expected mutations in that gene. Expectations were based on a 3-parameter random null-model of mutagenesis, where mutations in genes experience no selection (section A.1.1 on page 107). We generated expected histograms, by binomially sampling nonsynonymous and synonymous mutations from each gene using $E_{\text{nonsynonymous}}/(E_{\text{nonsynonymous}} + E_{\text{synonymous}})$ as the probability of a nonsynonymous mutations in the gene. The number of trials in each gene sampling was constrained to equal the number of observed mutations and trials with no synonymous mutations (undefined ω) were discarded from the null distribution—just as

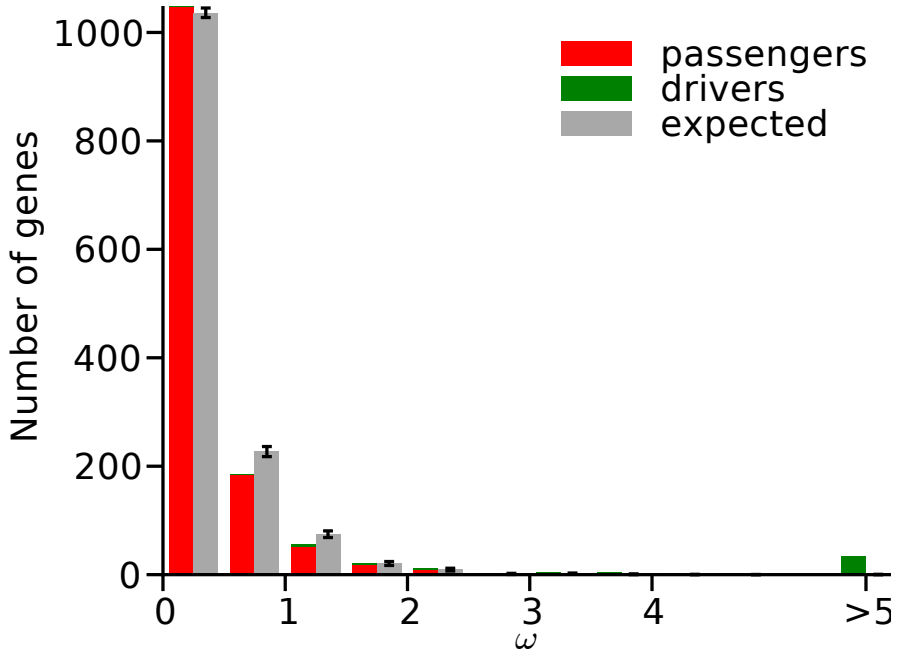


Figure 5.2: Cancer mutations show evidence of positive and negative selection. A histogram of the number of genes, under positive selection ($\omega > 1$) and negative selection ($\omega < 1$). Genes classified by COSMIC as passengers (red) generally experienced neutral or negative selection, while genes classified by COSMIC as putative drivers (green) generally experienced positive selection. Because of the very small number of observed mutations within each gene, there is considerable variance in the expected null-distribution of ω (grey) simply because of randomness. Indeed, most passenger genes used in the $\Delta PSIC$ analysis contained only 1 nonsynonymous mutation, which results in an undefined ω and could not be included in figure 5.2. Nevertheless, the observed distribution has an inordinate number of genes with $\omega < 1$ as well as an inordinate number of genes with $\omega > 1$. Additionally, annotated passenger genes are enriched in the set with $\omega < 1$ and annotated driver genes are enriched in the set with $\omega > 1$. This suggests that there exist both passenger genes under very mild negative selection as well as driver gene under strong positive selection.

genes with no observed synonymous mutations were discarded from the observed histogram.

This distribution reaffirmed COSMIC's driver and passenger classifications. Genes with $\omega < 1$ likely experience purifying selection and these genes were generally classified as passengers by COSMIC. Conversely, genes with $\omega > 1$ likely experience positive selection and were nearly all classified as drivers by COSMIC. Most importantly, the shape of this distribution corroborates the narrative of a few strong drivers over-

laid with copious passengers experiencing nearly undetectable negative selection that we observe in both our modeling and $\Delta PSIC$ analysis: A total of 94% of genes had an observed $\omega < 1$, and their occurrence was only very moderately enriched relative to our neutral model—on the fringe of statistical significance: $p = 0.012$—and not nearly as pronounced as the signal for drivers. Nevertheless, the true extent of purifying selection in cancer may be greater than suggested by this analysis of the COSMIC database. Consider that some publications do not report synonymous mutations and consider that putative passengers in figure 5.2 on the previous page may actually be an amalgam of genomic loci under negative selection mixed with loci under positive selection. These two complications would cause the observed degree of purifying selection to be less than the extent of purifying selection in reality.

The rare driver genes, with $\omega > 1$, often exhibited extreme nonsynonymous substitution rates vastly greater than expected from a neutral model of evolution: ω of KRAS, TP53, BRAF, and PTEN were all greater than 40. This supports the prediction of our model that the commonly accumulated drivers in cancer are of very large beneficial effect. While driver of moderate benefit may exist in the population and accumulate with frequencies slightly above the neutral rate, we find that the most studied drivers are very powerful.

Although this genomic analysis of passenger mutations focused on missense substitutions, our model is generalizable to all inheritable (epi)genetic alterations, including those that are present at low frequency in the cancer population. Indeed, the length distribution of somatic copy number alterations (SCNAs) in cancer suggests these alterations are under purifying selection as well[38]. Hence, the total load of accumulated deleterious passengers in cancer may be greater than that estimated from single nucleotide mutations detected in genome sequencing.

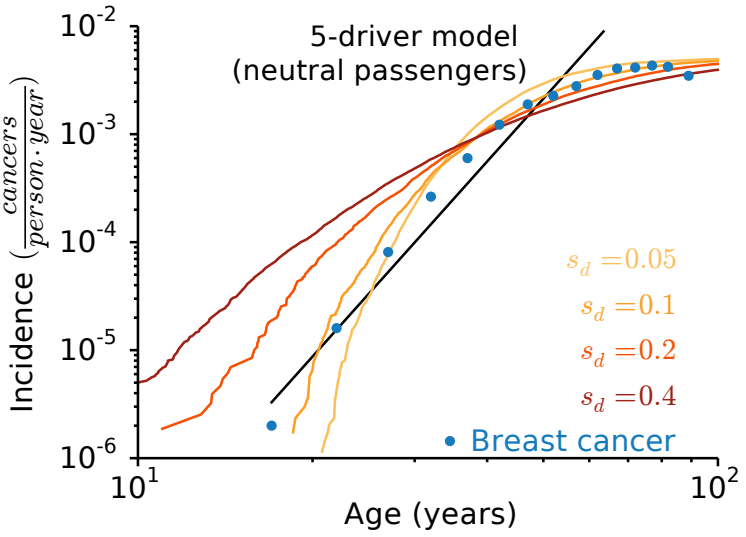


Figure 5.3: Predicted and observed breast cancer incidence rates verses age. Incidence rates in our model and the data both plateau at old age, but a traditional driver-only model ($I \propto t^k$) does not. Our model fits observed incidence rates best when s_d is large.

5.2 Evidence of a critical barrier in cancer age-incidence curves

Figure 5.3 presents the incidence rate of breast cancer versus age[55] alongside the predictions from our model and a classic driver-only model (section A.3 on page 110). The incidence rate was calculated by assuming that precancerous lesions arise with a constant rate r beginning at birth. Lesions then progress to cancer in time τ with probability $P(\tau)$, determined from simulations. The age-incidence rate $I(t)$ is then the convolution of $P(\tau)$ with the lesion initiation rate r. Since many lesions never progress and go extinct in our model, the incidence rate saturates at old-age: $I_{\max} = r \int_0^\infty P(\tau)\, d\tau = r P_\infty$, where P_∞ is the probability that a lesion ever progresses to cancer (see section A.2.1 on page 110 for details).

Observed age-incidence rates saturate with age, allowing us to roughly estimate the efficiency of cancer progression. We estimate that the rate of lesion formation in

breast cancer r is $\geq$ 10 per year, deducible in two ways: by multiplying the number of human breast epithelial stem cells by their rate of mutation into a lesion (section A.2.1 on page 110), or by considering the number of lesions observed in normal breasts[98]. By comparing this limit ($r \sim 10$ year^{-1}) to the maximum observed breast cancer incidence rate $I_{\max} \approx 10^{-2}$ cancers $\cdot$ year^{-1}, we find that $P_\infty \approx 10^{-3}$, or only $\sim$0.1% of lesions ever progress. Conversely, in a driver-only model (section A.3 on page 110), every lesion eventually progresses to cancer after sufficient time (i.e. $P_\infty = 1$) and incidence rates plateau only if the lesion formation rate is unrealistically low (0.01 per year). Good agreement between age-incidence data and our model is obtained for $s_d \approx 0.1 - 0.2$, and N^0/N^* is chosen such that $P_\infty = 10^{-3}$. This suggests that cancer begins at a population size far below N^*, where drivers are most often overpowered by passengers.

Likewise, figure 5.4 on the next page shows that most other cancer subtypes (84%) plateau at old age, indicating that inefficient progression is common. These findings are consistent with medical observations that very few lesions ever progress to cancer[13,118]. Like our model, these studies find that most cancers regress to undetectable size. Collectively, these results suggest that most cancers experience a barrier to progression that we believe is caused by passengers. Progression is inefficient and age-incidence curves are most consistent with our model for large s_d.

5.3 A wide and assymetric distribution of mutation totals suggests s_d is large

We looked at Somatic Nonsynonymous Mutations (SNMs) and Somatic Copy-Number Alterations (SCNAs) from over 700 individual cancer-normal sample pairs obtain

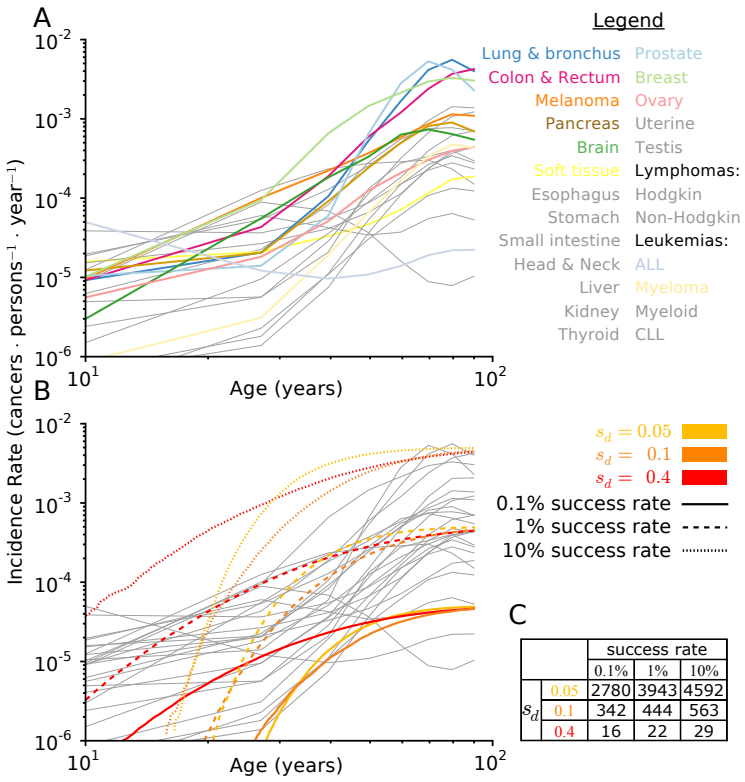


Figure 5.4: Age-incidence curves in our model depend primarily on s_d and match observed age-incidence curves well at mid/late life. **(A)** Incidence rate verses age for the 25 most common cancers in the SEERs database[55]. Nearly all cancers show incidence rates that rise rapidly at mid-life, but then plateau at old-age. Leukemias have flatter curves, suggesting that they need fewer drivers for carcinogenesis. Only colorectal cancer does not plateau. Instead, it exhibits a power-law relationship for all ages. Some incidence curves flatten at young ages, which has been linked to germ-line predispositions to cancer that expedite progression. Neither our model nor the traditional neutral-model of attempt to explain these childhood occurrences. **(B)** The predicted age-incidence curves derived from simulations match observed age-incidence curves in most cancer subtypes when proper parameters are chosen. The slope of predicted age-incidence curves is described by s_d: a larger s_d causes the slope of age-incidence curves to decrease. The location in the plateau of age-incidence curves is described by the success rate of cancer progression P_∞ multiplied by the lesion formation rate r. These two parameters introduce, essentially, a Gauge freedom into our comparison of simulations with theory. Hence, it is most useful to simply think of only one additional parameter (after s_d): a overall height of the plateau. This effective Gauge freedom is evident by the fact that the incidence curves with the same s_d look approximately the same, irregardless of their success rate. Only their overall height changes. For the predicted age-incidence curves plotted, $r = 5$. Of course this value, and the success rate, presumably vary considerably between cancer types. **(C)** The actual initial population size (N^0) needed to obtain various success rates of cancer progression from various s_d. Values were obtained by iteratively simulating various initial sizes until converging to an initial population size that led to the desired success rate. These values differed only mildly from the predictions of our analytical theory.

in breast[115], colon[21], lung[28], and skin[10] cancer (table 1.1 on page 4) to interrogate properties of drivers and passengers in these tumors that might support or refute our theory.

Analyzing SNMs and SCNAs separately and in aggregation yielded similar results (figure 5.8 on page 86, table A.1 on page 116). Figure 5.5 on the next page shows a wide and asymmetric distribution of the total number of SNMs in breast cancer. Our model predicts a similarly wide distribution of total SNMs due to the stochastic period of time that cancers linger at the critical population size N^* while accumulating mutations. To compare these data with various models of progression, we normalized the total number of mutations by their median (27), as several evolutionary parameters and sequencing decisions can alter these distributions by a multiplicative factor (section A.2 on page 108).

Our model agrees with the data when the effect size of drivers is large ($s_d \approx 0.4$). In contrast, a traditional 5-driver model (section A.3 on page 110, figure 5.6 on page 82), which neglects deleterious passengers (and, thus, a critical barrier), yields a narrower, less-skewed distribution than observed. The critical barrier in our model increases the variance and skew of the expected distribution because populations can linger around N^* for a highly variable period of evolutionary time (all while acquiring additional drivers and passengers). The traditional model of cancer presumes progression is more-or-less a methodical march towards cancer where every driver is an irreversible step closer to carcinogenesis. No declines in population size, due to passenger accumulation are possible. As such, the variance in progression time and resulting mutation totals are comparatively smaller.

Overall, our model agrees with the 11 most sequenced cancer subtypes with 100 or more sequenced tumors[72]. Our fit was best when $s_d \approx 0.1 - 0.6$ (figure 5.6 on the pre-

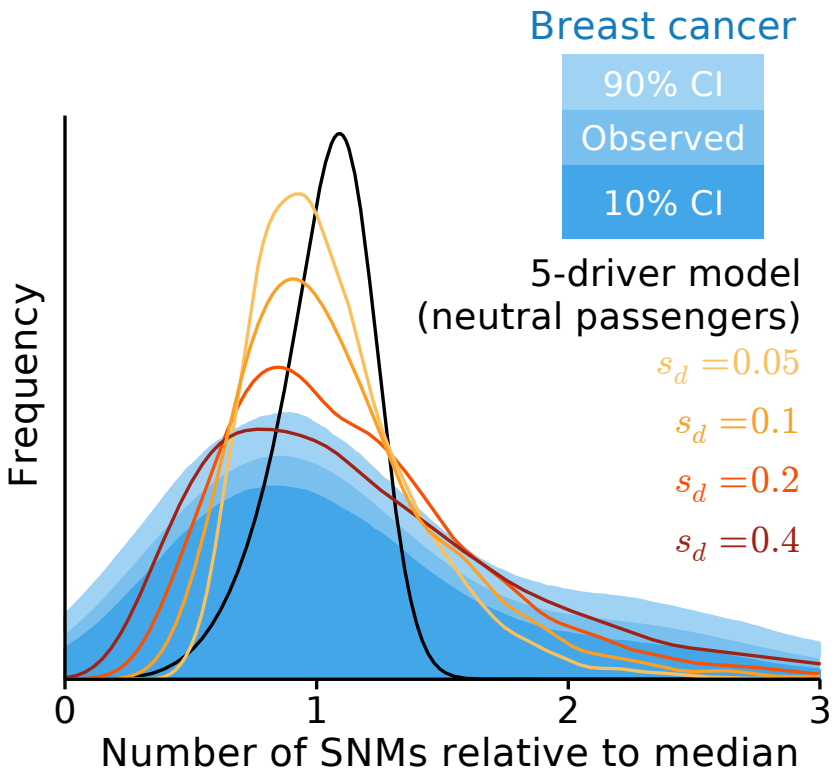


Figure 5.5: A wide and asymmetric distribution of mutation totals in breast cancer supports our model. Histogram (blue) of the collective number of protein-coding mutations (SNMs) in breast cancer, alongside predicted distributions. Our model, captures the width and asymmetry of the distribution when $s_d = 0.4$, while a driver-only model predicts a narrower, less-skewed distribution.

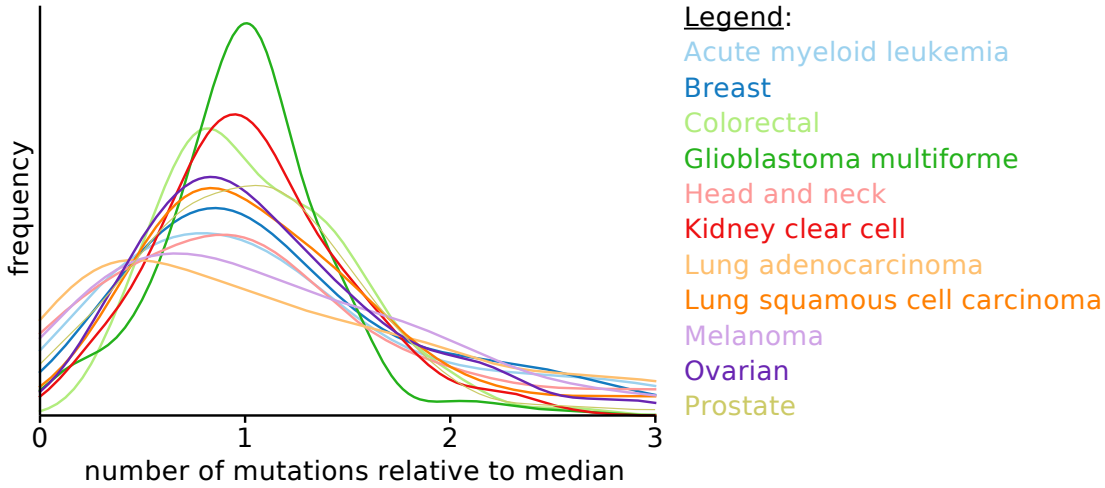


Figure 5.6: Distribution of mutation totals in cancers is highly dispersed and positively skewed. 11 cancer subtypes have 100 or more tumors sequenced via TCGA [72]. These subtypes all had widely-different distributions of mutation totals; however, the all appear to have a large degree of variance within their subtype and positive skew. Violin plots of each distribution are shown. In table 5.1 on the next page, we compared our model of cancer progression, for various s_d, alongside a traditional, neutral-passenger model of cancer progression (see section A.3 on page 110) to these distributions.

Table 5.1: Kolmogorov-Smirnov goodness of fit estimates of s_d (our model) and k (traditional model) from the most-sequenced cancer subtypes. We selected our best fitting model, and the best fitting traditional model (largest D statistic) using a Kolmogorov-Smirnov goodness of fit test and displayed the best fitting estimates of s_d or k alongside their quality of fit.

			Our model		Traditional model
Tissue	N	s_d†	p-value*	k‡	p-value*
Acute myeloid leukemia	132	0.8	0.802	1	0.072
Breast	120	0.6	0.860	1	0.045
Colorectal	230	0.2	0.956	2	0.011
Glioblastoma multiforme	219	0.1	0.070	3	0.187
Head and neck	178	0.8	0.120	1	0.077
Kidney clear cell	214	0.2	0.485	2	0.066
Lung adenocarcinoma	333	0.8	3.71×10^{-5}	1	4.35×10^{-7}
Lung squamous cell	178	0.4	0.752	1	1.10×10^{-3}
Melanoma	121	0.8	0.267	1	0.073
Ovarian	385	0.4	0.833	1	4.54×10^{-6}
Prostate	221	0.4	0.134	1	0.018

*2-sided probability that the observed and expected distributions are identical.
† Optimal theoretical fit to observed data for $s_d \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$
‡ See section A.3 on page 110 for our M-L Estimator of this quantity.

vious page). These estimates agree with $s_d = 0.16 - 0.58$ that were measured experimentally as growth rate changes of mouse intestinal stem cells upon mutations in p53, APC or k-RAS[119] (table 5.1). These experimental measurements and our estimates are also considerably larger than previous theoretical estimates of $s_d \approx 0.004$[15], where cancer progression was modeled as an exponential growth unaffected by passengers.

A driver-only model fits SNM distributions poorly and suggests that just 1-2 drivers are needed for cancer (table 5.1), which is inconsistent with known biology. Taken together, cancer genomics data and recent direct fitness measurements in mice[119] strongly support our model and refute the driver-only model.

5.4 Tug-of-war between drivers and passengers is manifested by their positive linear relationship in tumor samples

We then compared the number of drivers and the number of passengers observed in individual cancer samples to our model. Lesions that linger around N^* in our model acquire additional passengers *and* additional counterbalancing drivers, while lesions that progress quickly acquire fewer of each. This predicts a linear relationship: $n_d = s_p/s_d \cdot n_p + \text{constant}$ (chapter 4 on page 41), where the slope provides an estimate of s_p/s_d . We indeed observe a positive linear relationship between n_d and n_p in all tumors studied (figure 5.7 on the following page, table A.1 on page 116, $p < 0.08\text{–}10^{-6}$). Linearity was confirmed by regressing the aggregated data in log-log axes (figure 5.7 on the following page), which yielded $n_d \sim n_p^{0.98}$, consistent with $n_d \sim n_p$. Regressing n_d on n_p, and using 10^4 bootstrapped samples to estimate the confidence, we obtain an $s_p/s_d \approx 0.005\text{–}0.05$ (figure 5.7 on the next page). Using our estimate $s_d = 0.1\text{–}0.6$ we obtain damaging effect of a passenger mutation $s_p \approx 5 \cdot (10^{-4}\text{–}10^{-2})$. These estimates are consistent with the effects of germ-line SNMs in humans, where 64% of mutations decrease fitness by $10^{-5}\text{–}10^{-2}$ [14]. In summary, this analysis shows that passengers are indeed deleterious and $\sim 100\times$ weaker than drivers.

We considered and refuted several alternative explanations for the observed positive linear relationship between drivers and passengers. First, variation in the tumor stage or the rate/mechanism of mutagenesis cannot explain the observed relationship (table A.1 on page 116). Second, SCNAs and SNMs are uncorrelated, so the relationship could not results from their differing effect sizes (table A.1 on page 116, figure 5.8 on page 86). As the data disagrees with these alternate hypotheses, we conclude that cancer indeed proceeds as a tug-of-war between drivers and passengers.

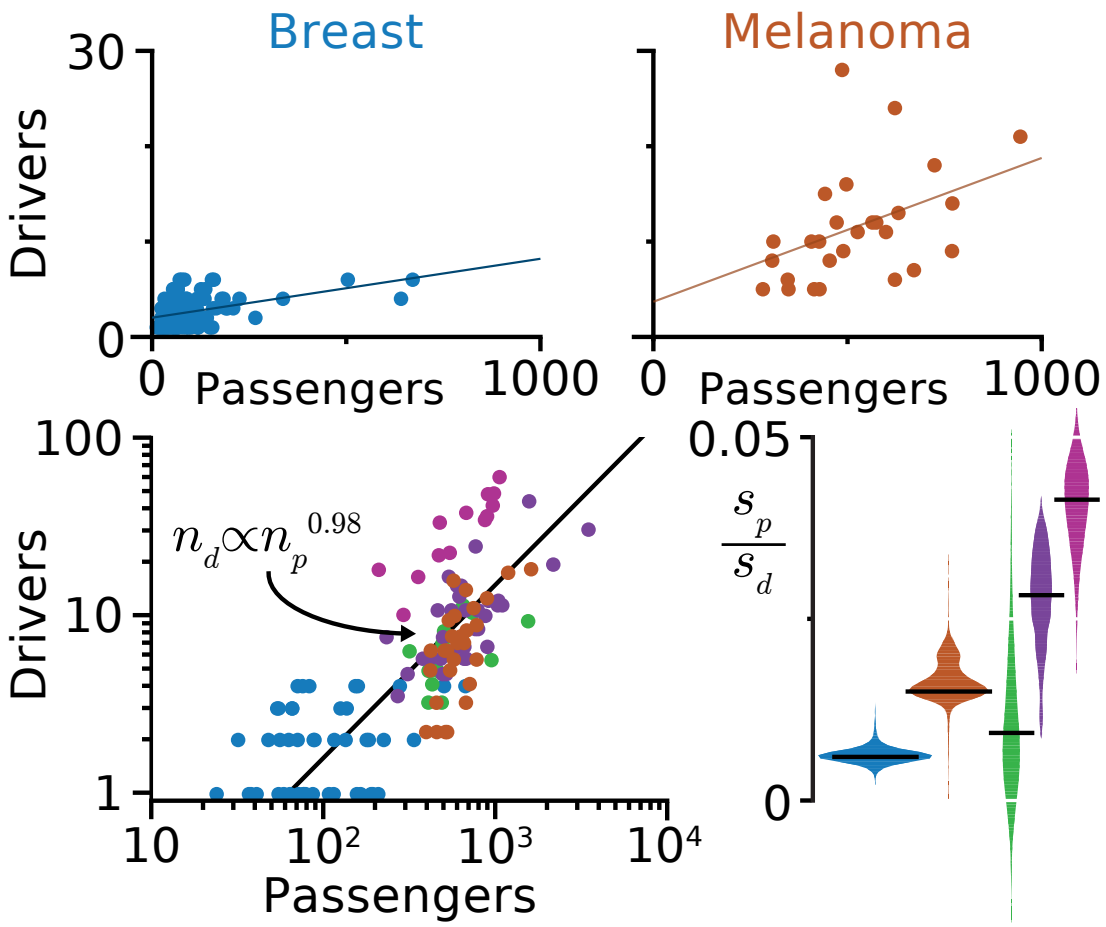


Figure 5.7: A positive linear relationship between drivers and passengers accross cancer subtypes. (Top) The total number of aggregate SCNA and SNM drivers verses the total number of passengers in sequenced tumors (points) fitted by linear regression. (Lower, left) Aggregated cancer genomics data plotted on log axes, with the y-intercept from each subtype's linear fit subtracted (Lung cancer (green), Colorectal cancer (MIN$^-$ dark purple, MIN$^+$ light purple)). (Lower, right) The distribution of slope values obtained by bootstrapping $10,000$ samples of each tumor type. All cancers exhibit positive slopes, ($p < 0.08 - 10^{-5}$), suggesting estimates of s_p/s_d of: Breast 0.0060 ± 0.0010, Melanoma 0.016 ± 0.003, Lung 0.0094 ± 0.0093, Colorectal MIN$^-$ 0.028 ± 0.007, and MIN$^+$ 0.041 ± 0.006.

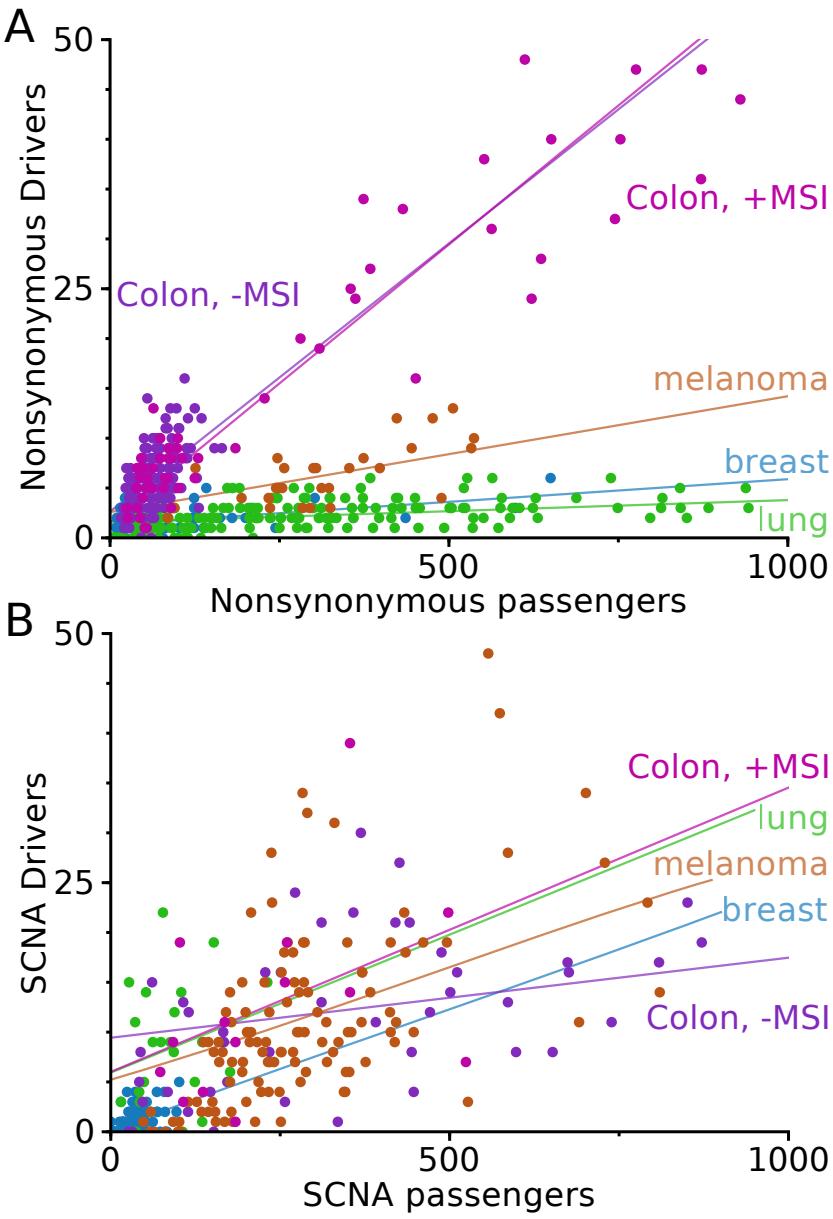


Figure 5.8: Somatic Nonsynonymous Mutations (SNMs) and Somatic Copy Number Alterations (SCNAs) exhibit similar positive linear relationships among cancer subtypes. **(A)** A positive linear relationship is observed between driver and passenger SNMs in all cancer subtypes studied here. This suggests that additional SNM passengers are being counterbalanced by additional drivers, and is consistent with our conclusions in the main text. Slope, y-intercepts, and the statistical significance of each best-fit line can be found in table A.1 on page 116. **(B)** Positive linear relationship is also observed in SCNAs. The similar slopes and y-intercepts of SCNAs to SNMs supports our assumption that SCNAs and SNMs can be aggregated in analysis and modeling.

Collectively, these results strongly support our theory that deleterious passengers accumulate in cancer, evade natural selection, and present a barrier to progression that lesions stochastically overcome. The data suggests that these dynamics are possible because drivers are rare strong-effect mutations and alterations, while passengers are common, yet mildly deleterious.

6

Clinical outcomes & passenger therapies

6.1 Accumulated Passenger Mutations Can Be Exploited for Cancer Treatment

Using our evolutionary model, we probed how cancers that accumulated passenger alterations respond to passenger-centric treatments. We tested two strategies: (i) increasing the overall mutation rate (μ), thus increasing the rate of passenger accumula-

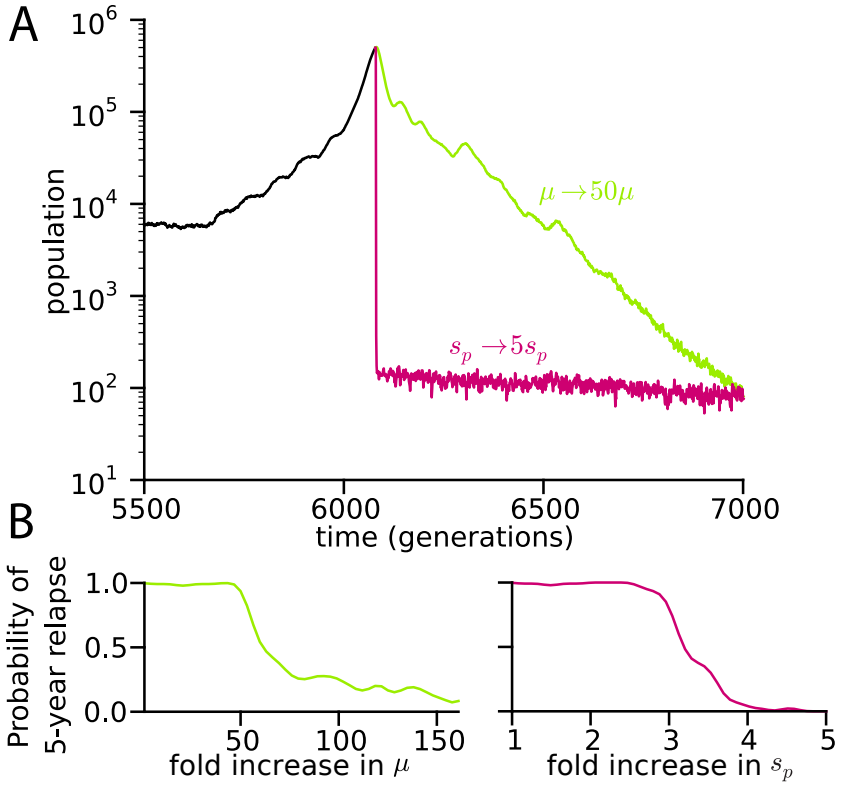


Figure 6.1: Deleterious passengers can be exploited for treatment. **(A)** Cancers grown to 10^6 cells are treated by increasing the mutation rate (green) or deleterious effect of passengers (magenta). Both strategies lead to reduction in cancer size. **(B)** Much smaller increase of the deleterious effect of passengers is sufficient to prevent 5-year relapse.

tion, and (ii) magnifying the deleterious effect of passengers (s_p), as described below. Figure 6.1 demonstrates that both strategies reduce cancer size; however, mutagenic strategies require more severe increases (50-fold) in the mutation rate to succeed (figure 6.1), whereas fivefold magnifications of passengers' deleterious effect suffice. Even with large mutation rate increases, the probability of 5-year relapse following treatment initiation is considerable (figure 6.1). This behavior resembles patient responses to existing chemotherapeutic agents that elevate mutation rates.

In practice, increasing the deleterious effect of passengers, both mutations and

chromosomal alterations[108], could be achieved by inhibiting cellular mechanisms that buffer against the effects of mutations or incorrect protein dosage[41]. Hence, the deleterious effect of passengers could be increased by targeting chaperones, proteasomes, or other components of UPR pathways[25]; or by elevating ER stress[26]; or by stimulating protein misfolding using hyperthermia[125]. These passenger-mediated therapies should specifically affect cancer cells because somatic mutations are generally rarer in normal tissues[22]. For example, a recent study of clonal mosaicism in human brains found only 1.5 SCNAs per adult sample[112], whereas a recent pan-cancer survey found 42 SCNAs per cancer[11].

Several experiments support this strategy of exacerbating passengers' effect. First, chaperones are widely expressed in cancer, indicative of poor prognosis[105], and their inhibition (or proteasome inhibition) exhibits antitumor activity[25]. Though other specific roles of chaperons and proteasomes in cancer were proposed, our framework suggests that cancers buffer against the effects of passenger alterations using UPR machinery. In our paradigm, inhibiting the UPR unleashes the effects of accumulated passengers. Recent discoveries that aneuploidy and chromosomal imbalance lead to proteotoxic stress[108] and a dependence on the UPR for survival[26].

6.2 The adaptive barrier and critical mutation rate explain cancer treatment outcomes

Chemotherapy and radiation are valued for their ability to kill rapidly dividing cells; however, our model shows that the elevation of mutation rates (including SCNAs and aneuploidy) by these therapies dramatically affects cancer survival. We use the phase diagrams from figure 4.1 on page 46 to rationalize outcomes of these and other treat-

ments.

In figure 6.2 on the following page, we present evolutionary paths of cancers—from hyperplasia, to cancer, to treatment, and relapse or remission—on top of the phase diagrams described above. Treatments succeed if they push cancer into the non-adaptive regime where the probability of growth is low, and fail if they do not. Our model suggests that chemotherapy succeeds, in part, because it moves cancers across the mutational threshold μ^*. Above this threshold, drivers seldom overpower the load of segregating passengers, making re-adaptation difficult. Driver-targeted therapies (that eliminate an oncogene's benefit) must bring $N < N^*$ to succeed.

Cancers with higher loads of mutations/alterations are closer to the critical mutation rate and should be most susceptible to mutagenic meltdown. Several recent studies[12,23,121] found that patient survival from breast (all subtypes) and ovarian cancer was greatest when tumors harbored exceptionally high levels of chromosomal alterations. These findings are paradoxical for all previous models of cancer[12], where a greater mutation rate always accelerates cancer, yet fully consistent with our model (figure 6.2 on the next page).

Treatments exploiting cancer's load of deleterious passengers remain unexplored. Figure 6.2 on the following page shows that a relatively mild 3-5 fold increase of the deleterious effect of passengers s_p causes complete remission. Increasing s_p is doubly effective because it exacerbates accumulated passengers *and* slows down future adaptation. Below we discuss possible treatment strategies that would increase s_p.

We find that simultaneously increasing cancer's mutation rate and the fitness cost of passengers is most effective: more effective than would be expected from adding together their individual effects (figure 6.3 on page 93). Hence, combinations of mutagenic chemotherapy alongside treatments elevating the cost of passengers may be

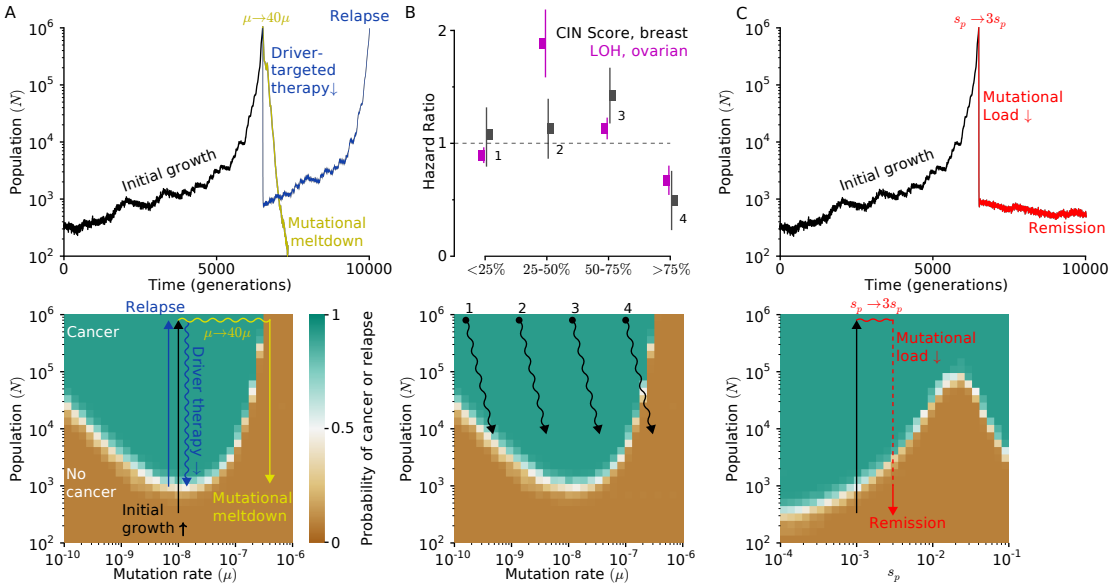


Figure 6.2: Mapping and interpreting treatment outcomes. **(A)** (Top) An adapted population (cancer) can be reverted to extinction by increasing the mutation rate (mutagenic chemotherapy) or by decreasing the population size (e.g. surgery or cytotoxic chemotherapy). However, relapse is possible. (Bottom) Our phase diagrams explain therapeutic outcomes: therapies that alter evolutionary parameters enough to push it outside of the adaptive regime cause continued population collapse; those that do not push a population outside of the adaptive regime re-adapt and relapse. **(B)** (Top) Cancers with intermediate mutational loads are the most aggressive [12,121], while patients with very high level of chromosomal instability are most effectively treated. (Bottom) This result is well explained by our phase diagrams, where cancer with very high mutation rates are susceptible to mutational meltdown. Yet this result is paradoxical for all previous evolutionary models of cancer. In this diagram, we assume traditional therapies decrease population size and mildy increase the mutation rate. **(C)** A relatively-mild three-fold increase in the effect of passenger mutations (s_p) leads to rapid population meltdown below N^*, thus without relapse.

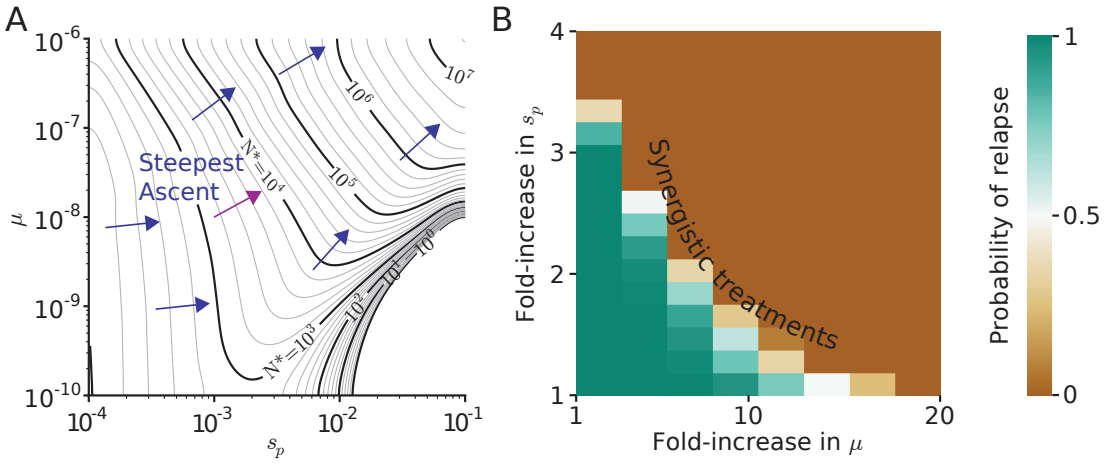


Figure 6.3: Combination treatments that increase mutation rate and selection against passengers work best. **(A)** Using the analytical theory describe in chapter 4 on page 41 and figure 4.1 on page 46, we plotted the critical population size N^* across evolutionary parameters as a contour plot. Optimal therapy, from an evolutionary perspective, should increase N^* along its gradient of steepest ascent (blue lines). From this 3-Dimensional perspective the interplay between μ and s_p is evident. For cancers with low mutation rates, only weak passengers (low s_p) can fixate. Thus, these cancers should be relatively immune to drugs that increase s_p. Cancers with high mutation rates fixate all passengers, making passenger-targeted therapies highly effective. At intermediate mutation rates, the most effective treatment would moderately increase both the mutation rate and s_p. **(B)** Via simulations, we tested our prediction that the gradient of steepest ascent is optimal for the magenta-colored vector in **A**. 50 cancers with $\mu = 10^{-8}$, $s_p = 0.001$ grown to 10^6 cells were treated with combinations of mutagenic and s_p increasing therapy. Indeed, moderate increases in both parameters were more effective than would be expected from the lone treatments, thus confirming our prediction. These results underscore the importance of combinatorial therapies and evolutionary modeling for cancer treatment.

most effective; these therapies should also synergize with driver-targeted therapies.

In [unadapted] condition there is no place for industry, because the fruit thereof is uncertain, and consequently, not culture of the earth, no navigation, nor the use of commodities that may be imported by sea, no commodious building, no instruments of moving and removing such things as require much force, no knowledge of the face of the earth, no account of time, no arts, no letters, no society, and which is worst of all, continual fear and danger of violent death, and the life of man, solitary, poor, nasty, brutish, and short.

Thomas Hobbes

7

Discussion and Future Directions

7.1 Conclusions

Cancer research has focused primarily on driver alterations with little attention to the overwhelming majority of potentially harmful passenger alterations that arise along the way. This "dark matter" of cancer genomes has not yet been previously explored in a systematic way. We developed an evolutionary model of cancer progression that

Table 7.1: The deleterious passenger model reproduces many properties of cancer. Many of the above phenomena would not be observed in our model without the inclusion of deleterious passengers. None of the above phenomena were pre-programmed into the model of neoplastic growth (i.e. population size was not fixed, nor was the number of mutations).

Phenomena Observed in our model	Experimental Observation
Clonal expansion, delayed growth, and extinction	2,54
Most lesions spontaneously regress	13,118
More mutations accumulate with high mutation rate	75
$\sim 50 - 300$ deleterious mutations accumulate	101,102
Significant heterogeneity in tumors cells, yet driver mutations are often clonal	74,126,19
Mutagenic therapies often relapse after a period of remission	110
The most mutagenic tumors respond best to traditional therapies	12,23,121

clearly demonstrates that deleterious passengers can accumulate in cancer, while our genomic analysis confirms that passengers presented in sequenced cancers have damaging phenotypes. Importantly, when cancer is modeled as balance between drivers and deleterious passengers, many observed phenomena in cancer are reproduced, including (i) slow initial and rapid late growth; (ii) a critical cancer size for dormancy or spontaneous regression; (iii) short-term response to mutagenic therapies; and (iv) substantial stochasticity of dynamics and (v) inefficiency in progression (table 7.1). These phenomena were not preprogrammed into the model, suggesting that the deleterious effect of passengers explains many properties of cancers.

In our evolutionary model of rapid adaptation, rare strongly-advantageous driver mutations accumulate amid frequent mildly-deleterious passenger mutations. In this tug-of-war process, populations either succeed and adapt, or fail and go extinct. Simulations and theory identify two regimes of dynamics: one where populations almost always adapt, and another where they almost always fail. We also found a critical mutation rate, above which populations quickly meltdown.

This general framework for adaptive asexual populations effectively characterizes the observed dynamics of cancer progression and therapeutic responses. We show that the late onset of cancer, evident in age-incidence curves, can be explained by a passenger-generated barrier to cancer, and does not require more complex models of cancer progression (e.g. a specific order of mutations, or multi-hit model, or variable mutation rate). We also considered a commonly used "two-hit model", where the first driver confers no fitness benefit, while two drivers confer a strong cumulative effect. This two-hit model also exhibits a barrier to cancer and behaves similarly to our model, yet with an effectively larger N^* and s_d.

Our framework suggests that most normal tissues reside in a regime where cancer progression is exceedingly rare. Most lesions fail to overcome the adaptive barrier at N^*. This implies that observed tumors acquired drivers faster than a mean trajectory, which may be testable via phylogenetic analysis.

Clinical cancers, on the contrary, reside above the adaptive barrier in a rapidly adapting state. Successful therapies must push a cancer below this adaptive barrier N^* to succeed. In our framework, this entails moving populations below N^* or increasing mutation rate above the critical value μ^*. A broad range of cancer data allowed us to thoroughly test the applicability of these barriers.

We tested our model and estimated its parameters using age-incidence curves, cancer exome sequences from ∼1,000 tumors in four cancer subtypes, and data on clinical outcomes. Age-incidence curves support our hypothesis that nearly all lesions fail to progress and allowed us to estimate the fitness benefit of a driver as $s_d \approx 0.1$–0.6 — in good agreement with direct experimental measurements[119].

We found that individual SNM passengers exhibit signatures of damaging mutations, and thus likely affect the fitness of cancerous cells. The length distribution

of SCNAs in cancer suggests that these alterations are also under purifying selection as well[38]. Deleterious passengers could result from any inheritable change in cell state, e.g. stable changes in cell signaling, or epigenetic modifications. Genomics data also affirmed that passengers are deleterious with an $s_p \approx \frac{1}{100} s_d$, yielding $s_p \approx 5 \cdot (10^{-4}\text{–}10^{-2})$, have a damaging effect 100-times smaller that the effect of driver, but are a hundred times more numerous than drivers. Taken together, these data support the notion of a tug-of-war between rare large-effect drivers, verses frequent mildly-deleterious passengers.

We focused on the evolution of cancer, but our model should generalize to other adaptive asexual processes. Consider a small population entering a new environment. Fluctuations in its size often lead to its extinction. Occasionally, however, the population may acquire several new highly-advantageous traits for this new environment, allowing it to rapidly expand its size and avert extinction. Both its evolutionary parameters[48] and behavior[9] mirror our model. Our mathematical framework further explains why these populations sometimes adapt, yet often fail.

7.2 Important considerations for future evolutionary models of cancer progression

Our simple model captured many of the phenomena of carcinogenesis and treatment without be preprogrammed to do so. As such, we expect the to be very instructive for a large array of future work. However, there are several evolutionary questions that I believe are particularly pressing given our recent findings.

7.2.1 Modeling the gain and/or loss of a mutator phenotype

We assumed that the mutation rate is constant in our evolving population. Often an enhanced mutation rate in cancer cells is a result of driver mutations that increase cell fitness by eliminating cell-cycle checkpoints, which also ensure that DNA is faithfully repaired. However, an increased mutation rate can also be the direct result of mutations in DNA repair genes[75]. These mutator phenotypes are believed to be most beneficial if they arise early[96], but, if mildly-deleterious passengers are considered, may become disadvantageous in later stages of cancer progression.

Our analysis agrees with the current belief that a mutator phenotype is most beneficial early in tumor progression. Early on, many driver mutations are accessible and the rate of progression is limited by the rate at which drivers mutate. Because mildly-deleterious passengers are not readily selected against, increases in the mutation rate should not increase their probability of fixation. However, later in tumor progression, a mutator phenotype should become disadvantageous to the population as a whole. The accumulation of drivers should become limited by clonal interference (section 4.1.4 on page 57), rather than being limited by accessibility of drivers and the mutation rate. With a larger population size, natural selection should be capable of weeding-out deleterious passengers, but only if the mutation rate is moderate. Hence, decreases in the mutation rate can decrease v_p when the population size is large. Lastly, as the population begins to adapt to its new microenvironment (evident by its increase in population size) the average benefit of new drivers should decline. Because, $\mu^* = s_d$ (section 4.2 on page 63), this should lead to a lower threshold for mutational meltdown.

Thus, decreases in the mutation rate in the later stages of cancer progression should

be beneficial for the population as a whole, if not necessary. Yet no thorough investigation of the selective pressures for or against a mutator phenotype in cancer have been performed. These forces should depend critically on the extent of deleterious passengers in the population. In natural populations, theoretical work suggests that complete genetic linkage (i.e. asexual population dynamics, as we model in cancer evolution) can result in mutator phenotypes being beneficial to individuals at nearly any mutation rate. In fact, the genetic linkage between mutator phenotypes and the beneficial drivers they may create can lead to runaway mutation rates that harm the total fitness of the population[42]. Whether this applies to cancer is unknown. If a runaway mutation rate can be fostered, or even selected for, then new therapeutic strategies may be possible.

7.2.2 Understanding the spatial structure of cancer

As with many evolutionary models, we assumed that our population was 'well-mixed', i.e. each individual was equally able to reproduce (genetics notwithstanding) at any given time, regardless of its position in space. In reality, tumors exist in a thee-dimensional environment, which experiences strong hydrostatic pressures, pH and oxygen gradients, contact inhibition, and paracrine signaling factors. In general, spatial segregation and environmental constraints have a tendency to undermine natural selection and increase the effects of genetic drift. Unlike some confounding factors in evolutionary theory, these changes generally cannot be subsumed into a alteration of the effective size, as spatial constraints affect each evolutionary metric differently. Thus, a careful evolutionary investigation of these spatial constraints would be very informative.

Spatial structure in tumors has been shown to prolong the waiting time to can-

cer[81]. As many precancerous populations do not progress to cancer within a human lifetime in well-mixed populations, the greater time constraint that spatial segregation imposes on cancer progression should encourage an even higher mutation rate, and necessitate that drivers be even more advantageous in cancer. How spatial structure affects the accumulation of mildly deleterious passengers remains unexplored. In all likelihood, spatial structure should continue to diminish the effectiveness of natural selection and increase the fixation probability of moderately deleterious passengers.

Spatial structure in tumors has also been shown to increase the degree of genetic diversity[45]. In general, the size of fluctuations of segregating alleles diminishes in spatially segregated populations. This increases their resident time in the population, and thus the number of segregating alleles at any one time. Understanding the degree of genetic diversity and how this diversity is modified by mutations with a mildly deleterious effect is critical to understanding the evolution of resistance in cancer.

7.3 Implications for cancer therapy and targeting deleterious passengers

Clinicians could exploit deleterious passengers by either increasing the mutation/alteration rate in cancers with already high mutation rates, or by increasing the deleterious effect of passengers. Clinical data indeed show that cancers with a higher load of chromosomal alterations, closer to μ^*, respond better to treatments[12,23,121]. PARP inhibitors, which increase DNA damage in BRCA1/2–positive tumors, may already be curing patients by inducing mutational meltdown[73].

7.3.1 Cancer's mutation rate as a biomarker for mutagenic therapy

A cancer's elevated mutation rate is a double-edged sword, it increases the likelihood of new drivers and mutations that might lead to drug resistance, but also introduce mutations that are harmful to the cancer cells. Our theory argues that clinicians should try to keep cancer's away from the optimal mutation rate μ^{opt}. Hence, patients with CIN or MIN might be best suited for mutagenic chemotherapies and radiation. Likewise, patients with a low or moderate mutation rate should not be treated with these therapies—they may make the tumor more aggressive. Fortuitously, these patients are, instead, ideal candidates for driver-targeted therapies, as their tumors are the least likely to evolve resistance.

We also argue that an increase in the mutation rate should increase the variation in fitness and phenotypes of a cancerous population. This theory can be tested experimentally, by comparing the morphology of cancer cells with a mutator phenotype to those without a mutator phenotype. We predict that cells with a mutator phenotype will exhibit greater morphological variability. This morphological variability may be exploitable by immunotherapies.

Morphological variability and genetic variability within a cancer population should be important biomarkers for inferring a patient's response to mutagenic therapies. Consider that a population's variation in fitness is proportional to μ_p (figure 3.6 on page 31)*. This implies that, if this standing genetic variation is large, then either the mutation rate is large, or the selection against passengers is small (and thus passengers are segregating in the population for a longer period of time). Tumors with a large degree of genetic variably, but low overall mutation rate can only exist in our

*This result also follows from Fischer's Fundamental Theory of Natural Selection and should hold even if many of the assumptions of our model are wrong

model if they have found a way to reduce the deleteriousness of passengers, perhaps by mutations in UPR pathways or the Major Histocompatibility Complex. Consequently, these tumors will theoretically be most resistant to mutagenic chemotherapy and thermotherapy, but most susceptible to chaperone and proteasome inhibitors.

7.3.2 Exacerbating passenger's deleterious effects

Passenger mutations and alterations can be deleterious by gain-of-function toxicity via proteotoxic/misfolding stress[108,59], or by eliciting an immune response to mutated epitopes[120]. Their damage to cancer cells could be magnified by (i) inhibiting unfolded protein response (UPR) pathways and proteasomes[108], (ii) hyperthermia that further destabilizes mutated proteins and clogs UPR pathways[103], or (iii) by activating an immune response[106]. Intriguingly, all these strategies are in clinical trials, yet none are believed to work by exacerbating passengers' deleterious effects. We offer an alternative explanation for their efficacy and suggest that these therapies will be most effective if used collectively and in cancers with many passengers.

One of the major limitations of driver-targeted therapies is that they can be defeated by a cancer's ability to rapidly evolve resistance by acquiring new mutations. Our approach, of increasing the deleterious effects of passengers, is different as it targets not only existing cancer cells, but also cancer's ability to accumulate new mutations—effectively decreasing its evolvability. Thus, these therapies should work well in combination with driver-targeted therapies.

It is important to consider that therapies that exacerbate deleterious passengers in cancer could also exacerbate the deleterious effects of mutations in the germ-line or somatic mutations outside of cancers. While there is evidence that these mutational loads are low[78], this possibility should be investigated further. Patients with

lower levels of protein-destabilizing germ-line mutations would be ideal candidates for passenger-targeted therapies. Like many cancer therapies, keeping the therapeutic drugs localized to the tumor site would be beneficial.

7.4 Deleterious passengers may explain metastatic inefficiency

The transformation of cancer from localized to metastatic is a highly lethal and poorly understood process. Evolutionary dynamics drive the primary tumor population, but also are important in the selection of successful metastases. The process of metastasis is highly inefficient, yet is the eventual endpoint of most cancers and usually results in patient death. While we yet lack the ability to predict where a given patient's cancer will metastasize, patterns of metastasis are observed and well documented for each different type of primary tumor. This suggests that while the process is random, it is driven by deterministic factors that we do not yet fully understand. The popular 'seed and soil' hypothesis states that not only are primary tumor— the'seed'—factors important, but the final stromal site of arrest—the 'soil'—also plays a critical role.

Metastasis is a highly inefficient process; estimates of hematogenous shedding of tumors cells are as high as 10^6 cells per gram of tissue[17], yet the incidence of clinically viable sites of metastasis does not reflect this number. There are many instances where patients will have evidence of tumor cells flowing in their blood without any evidence of any metastatic disease[89]. There is also evidence of patients with small, subclinical deposits of tumor cells in their marrow space (and assumedly other unassayed spaces) who never go on to develop overt clinical metastasis[35]. Indeed, in most cancers, these patients are not considered to have metastatic disease.

Deleterious passengers may explain inefficient metastatic inefficiency, just as they

have explained inefficient tumor progression. Indeed, metastases harbor more passengers than the primary tumor and experience a population bottleneck during dissemination that may push populations below the critical population size N^* that we identified in primary tumor progression. If passengers are partially responsible for metastatic inefficiency, then targeting them would have the added benefit of safeguarding against metastatic disease. Future models of this complicated progression should prove useful to clinicians.

A

Methodology

A.1 Analysis of somatic mutations in cancer for $\Delta PSIC$ analysis and for signatures of positive and negative selection

All cancer mutations were collected from the ongoing COSMIC database at http://www.sanger.ac.uk/genetics/CGP/cosmic/[36]. COSMIC and TCGA, along with other cancer genomics consortia, have focused on identifying driver mutations (i.e.

distinguishing drivers from passengers) by their recurrence in multiple patients or samples[36]. Using COSMIC, we identified 4,195 missense passenger mutations (non-synonymous, amino acid changes) from a total of 116,977 mutations. We defined a mutation as a passenger if it arose in a gene not listed in the census of possible cancer-causing genes[39]. These 4,195 'passenger' mutations show no recurrence and are dispersed across 3,172 genes, further supporting their classification as passengers. We then contrasted these mutations with driver mutations and three reference datasets:

1. benign, common human non-synonymous SNPs;
2. simulated de novo mutations (randomly generated using a cancer-specific 3-parameter model described in detail below); and
3. damaging, pathogenic missense mutations causing Mendelian human diseases (from the Human Gene Mutation Database, HGMD).

Common SNPs and disease causing mutations were obtained previously for validation of POLYPHEN2[1]. In our more stringent classification of passenger mutations, we discarded: 1) all passengers in genes that harbored more than one passenger, 2) passengers in any genes where $\omega > 1$ (figure 5.2 on page 75), and 3) passengers that were not 'confirmed' somatic mutations in the COSMIC dataset (only 29.4% of mutations in the database were 'confirmed' by follow-up Sanger sequencing). Mean $\Delta PSIC$ for this stringent set of passengers did not different significantly ($p < 0.69$) from our original set, so it was not used for further controls as it greatly reduced sample size.

To stratify passengers into various subsets, we used several resources. 372 passenger mutations were classified as 'Homozygous' by COSMIC, presumably due to some kind of Loss of Heterozygosity event. 'Housekeeping' genes, were 195 genes with passenger mutations and with one-to-one orthologs in S. cerevisiae, identified using

InParanoid[97]. These genes are well expressed in humans, so we believe it is highly likely that they are expressed in cancer. We could not directly normalize mutations in our dataset by their expression levels because mutations in the COSMIC database derive from varied literature sources (which often lack direct expression data). 881 non-COSMIC, non-synonymous passenger mutations were obtained from The Cancer Genome Atlas' analysis of 38 Multiple Myeloma genomes[22]. This subset was used as a control to ensure that any biases, which COSMIC might introduce via literature curation, did not account for our observed scores.

A.1.1 A pan-cancer null model of mutagenesis

To parameterize our random model of pan-cancer mutations, we collected all 1,128 synonymous mutations present in COSMIC at the time of this study. Given our sample size, we parameterized our model to account for 3 types of point mutations: transversions, CpG to TpG transitions, and all other transitions, as these 3 processes explained observed synonymous substitution patterns best (see below). Because some genes in COSMIC, like KRAS or TP53, are sequenced more often than others, we weighted both our estimated parameters and simulated mutations by the frequency with which each gene was sequenced [*]. We weighted genes in this manner because the frequency that genes were sequenced was highly variable and highly skewed, especially towards driver genes.

The null frequency of each mutually-exclusive mutation type f_i (CpG transitions,

[*]In the manifest of COSMIC, the curators note that they take care to record how often a gene is sequenced, since it is critical to determining whether-or-not a gene is abnormally mutated.

other transitions, & transversions) was estimated as follows:

$$f_i = \sum_{j}^{\text{all genes}} w_j \frac{O_{ij}}{P_{ij}} \tag{A.1}$$

Here, O_{ij} is the number of *observed* synonymous mutations in mutational class i, for gene j, while P_{ij} is the number of *possible* unique synonymous mutations for mutational class i of gene j, and w_j is the *weight* of gene j or fractional of cancer sequences reported in COSMIC belonging to gene j. This model explained the observed patterns of nonsynonymous mutations with greater Bayesian information criterion corrected log-likelihood than two-parameter models, as well as more sophisticated 10-parameter models or models developed for human germ-line mutations[114]. Using this model, random mutations were drawn with probability $f_i w_j$ from the set of all possible genome-wide, nonsynonymous mutations. These randomly-generated mutations were not only used as a null model for evolutionary conservation, but also as a neutral null-model to test for signatures of positive and negative selection in cancer genomes (figure 5.2 on page 75). Currently, there exist more sophisticated null-models of mutagenesis that leverage more data and account for patterns in mutagenesis[72].

A.2 Analysis of cancer genomes

To analyze mutations at a subtype-specific level (i.e. at the level of breast cancer, lung cancer, etc), we leveraged the sequencing and classifications methods used in the publication of each subtype's genomic dataset. We felt that a systematic meta-analysis was more appropriate at the subtype level, rather than pooling all data and developing our own analysis pipeline (as we did above). Mutation calling and classifications were done in the articles where the SNMs and SCNAs were first identified

in exome-sequenced tumor-normal pairs: breast[115], colon[21], lung[28], and skin[10] cancer (table 1.1 on page 4). MIN colorectal cancers were distinguished from non-MIN tumors in a previous study[21].

Driver genes for every subtype were identified using MutSig[72] (for potential NSM drivers) and GISTIC 2.0[86] (for potential SCNA drivers). To be defined as a 'driver', a mutation needed to arise in a gene with a Bonferroni-corrected enrichment p-value $\leq 5 \cdot 10^{-3}$. All other mutations were classified as 'passengers'.

We chose to normalize observed and expected distributions of mutation totals by their median (for both our model and the neutral passenger model). This was done because both models contain free parameters that arbitrarily alter their median. In the neutral passenger model, this parameter is $< n_p^{\max} >$, while in our model, T_p/T_d and s_p could be simultaneously adjusted to fit the observed median without altering other properties of the distribution (see equation (3.2) on page 34).

There were also 6 breast cancers (5% of the dataset) with mutation totals greater than 4 times the median. All models in our study poorly explain these cancers, so we excluded them from our analysis. The breast cancer distribution was then compared to the various expected distributions using a violin plot. Error bar violin curves denoting the bottom 5th and top 95th percentile of the observed distribution were generated by creating 10,000 violin plots from 10,000 bootstrapped resamplings of the observed distribution.

These analyses were repeated for the 11 cancer subtypes that currently have ≥ 100 sequenced exomes. Using a Kolmogorov-Smirnov test for quality-of-fit, we compared each observed distribution to our simulated distributions for various $s_d \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$ and identified the best-fitting value of s_d. We then repeated these quality-of-fit tests for the driver-only model and identified the values of k that

best explain the observed distributions (figure 5.6 on page 82, table 5.1 on page 83). In general, our model explains the observed distributions more accurately than the driver-only model, with $s_d \approx 0.1 - 0.6$, while the driver-only model leads to unrealistically small values of k (between 1 and 2). In fact, for many caner subtypes, our model explains the observed distribution to statistical resolution.

A.2.1 Inference of lesion formation rate

In the main text, we argue that r is at least 10 lesions $\cdot$ year^{-1} in breast epithelial. This lower-bound estimate was based on the assumption that r =(10^3 breast epithelial stem cells per mouse[63]) $\times$ ($2 \cdot 10^3$ human breast epithelial stem cells per mouse breast epithelial stem cell) $\times$ (10^{-5} initiating mutations per cell per year[78]) $\approx$ 20. Moreover, the number of lesions observed in normal breasts corroborates this estimate[98,123]. Thus, we eliminate the possibility that age-incidence curves can be explained by models which assume that most lesions eventually progress to cancer.

A.3 A traditional model of cancer progression with drivers and neutral passengers.

In the traditional model of cancer progression used to estimate age-incidence curves, it is assumed that a cancerous population transitions through k intermediate states before malignancy:

$$C_0 \xrightarrow{r_1} C_1 \xrightarrow{r_2} ... \xrightarrow{r_k} C_k$$

Simply put, these intermediate states and transitions correspond to the many phenotypic changes that occur within a tumor as it progresses[54]. The instantaneous

probabilities of each transition from one state to the next r_i can vary in the general case. Nevertheless, it has been shown that this predicts similar age-incidence rates to a model where transition rates are all the same[4]. Thus, for parsimony we only consider the case where all transition rates are the same constant r. Moreover, if the transition rates are drastically different from one another, then dynamics will largely be determined by the slowest rate alone. The faster rates are then no longer 'rate limiting steps' and can be neglected in timescale analysis.

From a genetic perspective, each transition corresponds to the acquisition of a new driver in the population. However from a mathematical perspective, this model is agnostic about the underlying molecular event that transitions a precancerous population from one state to the next. Thus, this model can be expand to include any set of heritable rate-limiting steps required for carcinogenesis: SNMs, SCNAs, alterations in DNA and histone moieties, stable changes in cell signaling cascades, etc. Therefore, we believe it is reasonable to assume that each rate-limiting step is the acquisition of a new driver, as has been presumed for many years[87].

We now consider the properties of this model when neutral passengers that do not alter progression also accumulate. The precancerous population is now defined by the state C_{n_d,n_p}. We consider the case where drivers accumulate at a fixed rate r_d and

passengers accumulate at different fixed rate r_p:

$$\begin{array}{ccccc} C_{0,0} & \xrightarrow{r_d} & C_{1,0} & \xrightarrow{r_d} & \cdots \\ \downarrow r_p & & \downarrow r_p & & \\ C_{0,1} & \xrightarrow{r_d} & C_{1,1} & \xrightarrow{r_d} & \cdots \\ \downarrow r_p & & \downarrow r_p & & \\ \vdots & & \vdots & \ddots & \\ & & & & C_{n_d,n_p} \end{array}$$

As before, cancer arises once enough drivers accumulate ($C_{n_d=k,n_p}$).

To interpret age-incidence data, as well as genomics data, we are interested in both the waiting time until cancer (t_{cancer}) and the total number of mutations ($n_p + k$). This model can be simplify by noting that there is a freedom in the units for which we measure time. In our simulations, time was measured in generations and then converted to years. Here, we chose to measure time in units of the driver transition probability r_d and will then convert this to years afterwards. Hence, $r_d = 1$ without loss of generality. Consider the quantity $\tau_{\text{cancer}} = t_{\text{cancer}} r_d$, as a dimensionless measure of the waiting time to cancer. It value will be roughly k on average. Because driver and passenger accumulation events are independent processes in this model, the joint probability of observing a cancer at time τ_{cancer} with n_p passenger mutations, $P(\tau_{\text{cancer}}, n_p | n_d = k, r_p)$, is:

$$P(\tau_{\text{cancer}}, n_p | n_d = k, r_p) = P(\tau_{\text{cancer}} | n_d = k) \cdot P(n_p | \tau_{\text{cancer}}, r_p) \tag{A.2}$$

This joint probability distribution provides a framework for identifying our quantities of interest.

The waiting times to cancer in this neutral-passenger model, has been previously shown to be a sum of exponentially-distributed waiting times[4], i.e. an Erlang or Gamma distribution, of the form:

$$\begin{aligned} P(\tau_{\text{cancer}}|n_d = k) = &\quad \text{Erlang}[\tau_{\text{cancer}}|n_d = k, r_d = 1] \\ = &\quad \frac{r_d^k \tau_{\text{cancer}}^{k-1} e^{-r_d \tau_{\text{cancer}}}}{(k-1)!} \\ = &\quad \tau_{\text{cancer}}^{k-1} e^{-\tau_{\text{cancer}}}/(k-1)! \\ \propto &\quad t_{\text{cancer}}^{k-1} \text{ , when } \tau_{\text{cancer}}/k \ll 1 \end{aligned} \tag{A.3}$$

Traditionally in this model, it is believed that very few precancerous population have enough time to progress, as lesion formation rates are much greater than cancer incidence rates. Hence, it is believed that age-incidence curves should be fit with only the beginning of this distribution: i.e. a power-law distribution (last line of equation (A.3)). We find that although this hypothesis explains age-incidence rates well at mid-age, it fails to explain the plateau in age-incidence rates seen at older ages in most cancer subtypes (figure 5.3 on page 77, figure 5.4 on page 79).

In this model, the total number of passengers accumulated is a Poisson distribution, if the time of progression t_{cancer} is known:

$$\begin{aligned} P(n_p|\tau_{\text{cancer}}, r_p) = &\quad \text{Poisson}[n_p| < n_p >= t_{\text{cancer}} r_p] \\ = &\quad e^{-<n_p>} < n_p >^{n_p} /n_p! \end{aligned} \tag{A.4}$$

Here, $< n_p >= t_{\text{cancer}} r_p$ is the mean number of expected passengers. The distribution takes this form because each passenger accumulation event occurs with an exponentially-distributed waiting time; a Poisson distribution describes the sum of events with exponentially-distributed waiting times in a fixed time interval. Because we do not know when a new lesion arrises, we must convolute this distribution with

our expected distribution of t_{cancer}.

The available time for cancer progression depends upon the length of a human life: t_{human}. If $t_{\text{human}} < t_{\text{cancer}}$, then the precancerous population will be unobserved in age-incidence and genomics data because the person died of an alternate cause prior to malignancy. Although the actual distribution of human lifetimes is complicated, we can still make inferences about the validity of this model by considering its extremes. Consider two opposing extreme cases: (1) when $t_{\text{human}} \gg \overline{t_{\text{cancer}}}$, all lesions eventually progress and are sequenced (i.e. a human lifetime is much greater than the mean time to cancer); and (2) when $t_{\text{human}} \ll \overline{t_{\text{cancer}}}$, only a few exceptional lesions progress (i.e. the mean time to cancer is much shorter than a human lifetime). We find that this first extreme predicts a much broader and more positively skewed distribution in the number of passengers, than the second case. In either case, both distributions exhibit similar predicted distributions in the number of total mutations. We used the second extreme where a human lifetime is much shorter than the mean time to progression, as this would be predicted by a power-law fit to age-incidence curves.

In the case where $t_{\text{human}} \gg \overline{t_{\text{cancer}}}$, accumulation of passengers follows a binomial process. Each accumulation event has probability $p = r_d/(r_d + r_p)$ of being a driver and probability $(1 - p)$ of being a passenger. Because the population has infinite time to progress to cancer, the binomial process continues until $n_d = k$ drivers accumulate. A binomial process that continues until k successes (i.e. drivers), will have a total number of failures (i.e. passengers) that samples a negative binomial distribution:

$$P(n_p|p,k) = \binom{n_p + k - 1}{n_p}(1-p)^{n_p}p^k \tag{A.5}$$

A negative binomial distribution with $p \ll 1$ (i.e. passengers greatly outnumber

drivers–as is the case in observed) reduces to a Poisson distribution.

In the case where $t_{\text{human}} \ll \overline{t_{\text{cancer}}}$, the waiting time to cancer follows a power law distribution (equation (A.3) on page 113). This, convoluted with the distribution of passengers expected for a particular t_{cancer} (equation (A.4) on page 113) yields the expected distribution of passengers for a cancer subtype:

$$\begin{aligned} P(n_p|k) = & \int_0^{\tau_h} P(\tau_{\text{cancer}}|k) P(n_p|\tau_{\text{cancer}}, r_p) d\tau_{\text{cancer}} \\ \approx & \int_0^{\tau_h} \frac{\tau_{\text{cancer}}^{k-1} k}{\tau_h^k} \frac{e^{-\tau_p} \tau_p^{n_p}}{n_p!} d\tau_{\text{cancer}} \\ \approx & 1/[*n_p!] \int_0^{\tau_h} e^{-\tau_p} \tau_{\text{cancer}}^{k-1} \tau_p^{n_p} d\tau_{\text{cancer}} \\ \approx & k/[\tau_h^k n_p! r_p^k] \int_{\tau_p=0}^{\tau_p=\tau_h r_p} e^{-\tau_p} \tau_p^{k-1+n_p} d\tau_p \\ \approx & \binom{n_p+k-1}{n_p} k! < n_p^{\max} >^{-k} \gamma(k+n_p, < n_p >^{\max}) \end{aligned} \tag{A.6}$$

Where $\gamma(s,x)$ is the normalized incomplete gamma function defined previously (equation (4.4) on page 48). In the above derivation, we eliminated a parameter by considering the quantity: $< n_p >^{\max} = \tau_h r_p$, which corresponds to the mean number of passengers expected for a person who lives until the maximum allowable time, τ_h. Obviously, for both predicted passenger distributions (equation (A.6) and equation (A.5) on the previous page) the total number of mutations $n_d + n_p$ is the expected number of passengers $P(n_p|k)$ *plus* the number of drivers k, which is constant.

A.4 Analysis of driver and passenger mutation covariates

To investigate possible alternate explanations for our observed correlation between drivers and passengers we interrogated a wide-variety of alternative covariates in table A.1 on the following page.

We observe a linear relationship between drivers and passengers that was predicted

Table A.1: Linear relationship between drivers and passengers cannot be explained by other tumor properties.

Cancer	Pearson's r	p-value*	N†	Spearman's ρ	slope‡	y-intercept
Drivers verses Passengers						
breast	0.423	$\mathbf{< 10^{-4}}$	100	0.413	0.006	2.02
lung	0.368	0.08	24	0.998	0.005	8.63
colon, MIN$^-$	0.624	$\mathbf{< 10^{-4}}$	49	0.985	0.009	17.11
colon, MIN$^+$	0.916	$\mathbf{< 10^{-5}}$	14	0.999	0.047	6.50
melanoma	0.749	$\mathbf{< 10^{-5}}$	29	0.995	0.015	3.69
All	0.937	$\mathbf{< 10^{-99}}$	217	0.992	0.042	-3.81
SNM drivers versus SNM passengers						
breast	0.390	$\mathbf{< 10^{-4}}$	100	0.178	0.005	1.34
lung	0.587	$\mathbf{< 10^{-17}}$	183	0.579	0.002	1.56
colon, MIN$^-$	0.990	$\mathbf{< 10^{-134}}$	159	0.569	0.054	2.65
colon, MIN$^+$	0.994	$\mathbf{< 10^{-60}}$	64	0.918	0.056	1.56
melanoma	0.878	$\mathbf{< 10^{-9}}$	29	0.974	0.012	2.59
All	0.924	$\mathbf{< 10^{-223}}$	536	0.592	0.050	-4.45
SCNA drivers verses SCNA passengers						
breast	0.443	$\mathbf{< 10^{-5}}$	100	0.433	0.024	0.17
lung	0.253	0.23	24	0.998	0.028	5.94
colon, MIN$^-$	0.770	$\mathbf{< 10^{-9}}$	49	0.984	0.008	9.44
colon, MIN$^+$	0.424	0.13	14	0.994	0.029	6.01
melanoma	0.559	$\mathbf{< 10^{-10}}$	121	0.663	0.023	5.23
All	0.573	$\mathbf{< 10^{-27}}$	309	0.962	0.012	5.76
SNMs versus SCNAs						
breast	0.052	0.61	100	0.149	0.237	64
lung	0.169	0.43	24	-0.548	1.268	334
colon, MIN$^-$	-0.080	0.58	49	-0.068	-0.021	137
colon, MIN$^+$	-0.265	0.36	14	0.045	-0.981	838
melanoma	-0.114	0.56	29	0.176	-0.183	431
All	0.331	$\mathbf{< 10^{-6}}$	217	-0.089	0.631	133
Drivers verses Pathological Grade						
breast	0.163	0.10	100	0.113	0.067	2.25
lung	-0.048	0.83	22	0.024	0.006	2.13
colon, MIN$^-$	-0.187	0.20	48	0.072	0.012	2.67
colon, MIN$^+$	-0.121	0.68	14	-0.338	0.004	3.09
melanoma	0.221	0.35	20	0.120	0.025	1.83
All	0.018	0.80	204	0.054	0.001	2.37
SNMs versus Pathological Grade						
breast	0.217	**0.03**	100	0.444	0.001	2.33
lung	0.193	**0.02**	158	0.235	0.000	1.79
colon, MIN$^-$	-0.084	0.30	156	0.039	0.000	2.51
colon, MIN$^+$	-0.045	0.73	63	-0.023	0.000	2.50
melanoma	0.119	0.62	20	0.114	0.000	2.06
All	-0.012	0.80	497	0.147	0.000	2.28
SCNAs versus Pathological Grade						
breast	0.248	**0.01**	100	0.235	0.007	2.18
lung	0.170	0.45	22	0.026	0.002	1.85
colon, MIN$^-$	-0.166	0.26	48	0.081	0.000	2.48
colon, MIN$^+$	-0.054	0.85	14	-0.333	0.000	2.99
melanoma	0.109	0.31	88	-0.253	0.000	2.06
All	-0.067	0.27	272	-0.112	0.000	2.38

Negative values are in gray.
*Statistically significant ($p < 0.05$) relationship are in bold
† Number of samples compared
‡ Denotes s_p/s_d when comparing drivers to passengers

by our model. Above the thick black line are relationships that appear to robustly covary, while the bottom half contains relationships that we believe are insignificant. In our model, driver's and passenger's linear relationship results from their competing effect: additional deleterious passengers must be overcome by additional drivers. However, this relationship could conceivably be explained by alternate factors; in particular, we were concerned that the mutation type, mutation rate, or aggressiveness of the tumor could also explain the observed relationship. The data above suggest that these possibilities are unlikely, thus supporting our conclusion that drivers compete with passengers.

Our rational for the above competing hypotheses and why we reject them are as follows:

1. SCNAs and SNMs might have drastically different effects on cancer progression and undermine our model. The slope and y-intercept between drivers and passengers is *approximately* equal in SCNAs and SNMs, suggesting the relative fitness effects of these mutations is similar.

2. Some cancers might progress via CIN, while others progress via an elevated point mutation rate. If so, a negative correlation between SCNAs and SNMs within tumor subtypes would be expected, which has been observed previously in a pan-cancer study[23] and within the aggregate colorectal dataset. However, this does not appear to be so in other tumor types, nor in colorectal cancer after segregation according to MIN phenotype. Thus, the observed patterns are not explicable by varying mutational mechanisms.

3. The relationship between drivers and passengers might be a result of variation in mutation rate. Variation in the mutation rate should only alter the wait-

ing time to cancer in the neutral-passenger model, and not alter mutation totals. Nevertheless, if variation in the mutation rate could explain the correlation between drivers and passengers, then stratifying tumors by their mutation rate should reduce the correlation. Because the relationship between drivers and passengers is persistent and strong within the MIN^+ and MIN^-subtypes—expected to have and not-have a mutator phenotype—we reject this hypothesis.

4. Tumors with more drivers and passengers might simply be more evolutionarily advance. Suppose some cancers are detected and sequenced later than others. These late cancers would not only possess additional drivers, but also additional passengers, even if passengers were neutral; thus, retaining the correlation between drivers and passengers. However, late-detected tumors with additional drivers should also be more advanced and more aggressive. We find that a tumor's pathological grade is uncorrelated with the number of drivers, refuting this possibility. Pathological grade was quantified by converting roman numerals into a linear scale (i.e. A Stage IV tumor corresponds to an aggressiveness of 4). Many tumors had intermediate grades that were given corresponding fractional values (e.g. a Stage IIIa tumor was translated into a 3.0, a Stage IIIb was given 3.3, and a Stage IIIc was given 3.7). Because this quantification of tumor grade may distort the scale of aggressiveness, Spearman's Rank correlations are provided.

For completeness, the relationship between SNMs and Pathological Grade,& SCNAs and Pathological Grade are also shown.

References

[1] Adzhubei, I. a., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4), 248–9.

[2] Aguirre-Ghiso, J. A. (2006). The problem of cancer dormancy: understanding the basic mechanisms and identifying therapeutic opportunities. *Cell Cycle*, 5(16), 1740–3.

[3] Andersen, M. H. & Schrama, D. (2006). Cytotoxic T Cells. *J Invest Dermatol*, 126(April 2005).

[4] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*, 8(1), 1–12.

[5] Bachtrog, D. & Gordo, I. (2004). Adaptive evolution of asexual populations under Muller's ratchet. *Evolution*, 58(7), 1403–13.

[6] Bauer, B., Siebert, R., & Traulsen, A. (2014). Cancer initiation with epistatic interactions between driver and passenger mutations. *J Theor Biol*, (pp. 1–9).

[7] Beckman, R. a. & Loeb, L. a. (2005). Negative clonal selection in tumor evolution. *Genetics*, 171(4), 2123–31.

[8] Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., & Nowak, M. a. (2007). Genetic progression and the waiting time to cancer. *PLoS Comput Biol*, 3(11), e225.

[9] Bell, G. & Gonzalez, A. (2009). Evolutionary rescue can prevent extinction following environmental change. *Ecol Lett*, 12(9), 942–8.

[10] Berger, M. F., Hodis, E., Heffernan, T. P., Deribe, Y. L., Lawrence, M. S., Protopopov, A., Ivanova, E., Watson, I. R., Nickerson, E., Ghosh, P., Zhang, H., Zeid, R., Ren, X., Cibulskis, K., Sivachenko, A. Y., Wagle, N., Sucker, A., Sougnez, C., Onofrio, R., Ambrogio, L., Auclair, D., Fennell, T., Carter, S. L., Drier, Y., Stojanov, P., Singer, M. a., Voet, D., Jing, R., Saksena, G., Barretina,

J., Ramos, A. H., Pugh, T. J., Stransky, N., Parkin, M., Winckler, W., Mahan, S., Ardlie, K., Baldwin, J., Wargo, J., Schadendorf, D., Meyerson, M., Gabriel, S. B., Golub, T. R., Wagner, S. N., Lander, E. S., Getz, G., Chin, L., & Garraway, L. a. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*, 485(7399), 502–6.

[11] Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. a., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. a., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. a., Janne, P. a., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. a., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., & Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283), 899–905.

[12] Birkbak, N. J., Eklund, A. C., Li, Q., McClelland, S. E., Endesfelder, D., Tan, P., Tan, I. B., Richardson, A. L., Szallasi, Z., & Swanton, C. (2011). Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Res*, 71(10), 3447–52.

[13] Bota, S., Auliac, J. B., Paris, C., Métayer, J., Sesboüé, R., Nouvet, G., & Thiberville, L. (2001). Follow-up of bronchial precancerous lesions and carcinoma in situ using fluorescence endoscopy. *Am J Respir Crit Care Med*, 164(9), 1688–93.

[14] Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5), e1000083.

[15] Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., & Nowak, M. a. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA*, 107(43), 18545–50.

[16] Brunet, E., Rouzine, I. M., & Wilke, C. O. (2008). The stochastic edge in adaptive evolution. *Genetics*, 179(1), 603–20.

[17] Butler, T. P. & Gullino, P. M. (1975). Quantitation of cell shedding into efferent blood of mammary adenocarcinoma. *Cancer Research*, 35(3), 512–516.

[18] Cameron, T. C., O'Sullivan, D., Reynolds, A., Piertney, S. B., & Benton, T. G. (2013). Eco-evolutionary dynamics in response to selection on life-history. *Ecol Lett*, 16(6), 754–63.

[19] Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. a., Green, A. R., Futreal, P. A., & Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA*, 105(35), 13081–6.

[20] Camps, M., Herman, A., Loh, E., & Loeb, L. a. (2007). Genetic constraints on protein evolution. *Crit Rev Biochem Mol Biol*, 42(5), 313–26.

[21] Cancer, T. & Atlas, G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330–7.

[22] Chapman, M. a., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, a. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. a., Meyerson, M., Lander, E. S., Getz, G., & Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), 467–72.

[23] Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*, 45(10), 1127–1133.

[24] Cole, A. M., Myant, K., Reed, K. R., Ridgway, R. a., Athineos, D., Van den Brink, G. R., Muncan, V., Clevers, H., Clarke, A. R., Sicinski, P., & Sansom, O. J. (2010). Cyclin D2-cyclin-dependent kinase 4/6 is required for efficient proliferation and tumorigenesis following Apc loss. *Cancer Res*, 70(20), 8149–58.

[25] Dai, C., Whitesell, L., Rogers, A. B., & Lindquist, S. (2007). Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis. *Cell*, 130(6), 1005–18.

[26] De Raedt, T., Walton, Z., Yecies, J. L., Li, D., Chen, Y., Malone, C. F., Maertens, O., Jeong, S. M., Bronson, R. T., Lebleu, V., Kalluri, R., Normant,

E., Haigis, M. C., Manning, B. D., Wong, K.-K., Macleod, K. F., & Cichowski, K. (2011). Exploiting cancer cell vulnerabilities to develop a combination therapy for ras-driven tumors. *Cancer cell*, 20(3), 400–13.

[27] Desai, M. M., Nicolaisen, L. E., Walczak, A. M., & Plotkin, J. B. (2012). The structure of allelic diversity in the presence of purifying selection. *Theor Popul Biol*, 81(2), 144–57.

[28] Ding, L., Getz, G., Wheeler, D. a., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. a., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. a., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. a., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. a., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. a., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. a., Meyerson, M., & Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216), 1069–75.

[29] Domazet-Lošo, T. & Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC biology*, 8(1), 66.

[30] Domingues, J. S. (2012). Gompertz Model : Resolution and Analysis for Tumors. *J Math Model Appl*, 1(7), 70–77.

[31] Eisenberg, E. & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TIG*, 29(10), 569–74.

[32] Elgendy, M., Sheridan, C., Brumatti, G., & Martin, S. J. (2011). Oncogenic Ras-induced expression of Noxa and Beclin-1 promotes autophagic cell death and limits clonogenic survival. *Molecular cell*, 42(1), 23–35.

[33] Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, 78(2), 737–756.

[34] Fischer, A., Greenman, C., & Mustonen, V. (2011). Germline fitness-based scoring of cancer mutations. *Genetics*, 188(2), 383–93.

[35] Flatmark, K., Borgen, E., Nesland, J., Rasmussen, H., Johannessen, H., Bukholm, I., Rosales, R., Hårklau, L., Jacobsen, H., Sandstad, B., et al. (2011). Disseminated tumour cells as a prognostic biomarker in colorectal cancer. *British journal of cancer*, 104(9), 1434–1439.

[36] Forbes, S. a., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, a., Teague, J. W., Futreal, P. a., & Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*, Chapter 10(March), Unit 10.11.

[37] Frank, S. A. (2007). *Dynamics of cancer: Incidence, Inheritance, and Evolution*.

[38] Fudenberg, G., Getz, G., Meyerson, M., & Mirny, L. a. (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol*, 29(12), 1109–13.

[39] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer*, 4(3), 177–83.

[40] Gabriel, W., Lynch, M., & Burger, R. (1993). Muller's ratchet and mutational meltdowns. *Evolution*, (pp. 1744–1757).

[41] Geiler-Samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M., Hartl, D. L., & Drummond, D. A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci USA*, 108(2), 680–5.

[42] Gerrish, P. & Colato, A. (2007). Complete genetic linkage can subvert natural selection. *Proceedings of the …*, 104(15).

[43] Gibson, M. a. & Bruck, J. (2000). Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J Phys Chem A*, 104(9), 1876–1889.

[44] Gillespie, J. H. (2004). *Population Genetics: A Concise Guide*. Johns Hopkins University Press.

[45] González-García, I., Solé, R. V., & Costa, J. (2002). Metapopulation dynamics and spatial heterogeneity in cancer. *Proc Natl Acad Sci USA*, 99(20), 13085–9.

[46] Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O., & Desai, M. M. (2012). Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc Natl Acad Sci USA*, 109(13), 4950–5.

[47] Gordo, I. & Charlesworth, B. (2000). The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics*, 154(3), 1379–87.

[48] Goyal, S., Balick, D. J., Jerison, E. R., Neher, R. a., Shraiman, B. I., & Desai, M. M. (2012). Dynamic mutation-selection balance as an evolutionary attractor. *Genetics*, 191(4), 1309–19.

[49] Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., & Lander, E. S. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4), 633–644.

[50] Guthrie, J. (1966). Effects of a synthetic antigonadotrophin (2-amino, 5-nitrothiazole) on growth of experimental testicular teratomas. *British journal of cancer*, 20(3), 582.

[51] Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, 445(7123), 82–5.

[52] Haeno, H., Iwasa, Y., & Michor, F. (2007). The evolution of two mutations during clonal expansion. *Genetics*, 177(4), 2209–21.

[53] Haigh, J. (1978). The accumulation of deleterious genes in a population-Muller's ratchet. *Theor Popul Biol*, 267, 251–267.

[54] Hanahan, D. & Weinberg, R. a. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646–74.

[55] Howlader, N., Noone, A., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Cho, H., Mariotto, A., Lewis, D., Chen, H., Feuer, E., & Cronin, K. (2013). SEER Cancer Stastistics Review 1975-2010.

[56] Iwasa, Y., Nowak, M. a., & Michor, F. (2006). Evolution of resistance during clonal expansion. *Genetics*, 172(4), 2557–66.

[57] Johnson, T. & Barton, N. H. (2002). The effect of deleterious alleles on adaptation in asexual populations. *Genetics*, 162(1), 395–411.

[58] Johnston, M. D., Edwards, C. M., Bodmer, W. F., Maini, P. K., & Chapman, S. J. (2007). Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proc Natl Acad Sci USA*, 104(10), 4008–13.

[59] Jordan, D. M., Ramensky, V. E., & Sunyaev, S. R. (2010). Human allelic variation: perspective from protein function, structure, and evolution. *Curr Opin Struct Biol*, 20(3), 342–50.

[60] Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J. G., Fidler, S., Weber, J., Pybus, O. G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P. R., Walker, B. D., Takiguchi, M., & Goulder, P. (2009). Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238), 641–5.

[61] Keightley, P. D., Kryukov, G. V., Sunyaev, S., Halligan, D. L., & Gaffney, D. J. (2005). Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res*, 15(10), 1373–8.

[62] Kim, S., Chin, K., Gray, J. W., & Bishop, J. M. (2004). A screen for genes that suppress loss of contact inhibition: identification of ing4 as a candidate tumor suppressor gene in human cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46), 16251–16256.

[63] Kordon, E. C. & Smith, G. H. (1998). An entire functional mammary gland may comprise the progeny from a single cell. *Development*, 125(10), 1921–30.

[64] Korolev, K. S. (2012). Time to extinction without drivers: Theory on driver passenger model. (4).

[65] Korolev, K. S., Avlund, M., Hallatschek, O., & Nelson, D. R. (2010). Genetic demixing and evolution in linear stepping stone models. *Rev Mod Phys*, 82(2), 1691–1718.

[66] Korolev, K. S., Müller, M. J. I., Karahan, N., Murray, A. W., Hallatschek, O., & Nelson, D. R. (2012). Selective sweeps in growing microbial colonies. *Phys Biol*, 9(2), 026008.

[67] Korolev, K. S., Xavier, J. B., & Gore, J. (2014). Turning ecology and evolution against cancer. *Nat Rev Cancer*, 14(5), 371–380.

[68] Kryukov, G. V., Schmidt, S., & Sunyaev, S. (2005). Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet*, 14(15), 2221–9.

[69] Kumar, M. S., Pester, R. E., Chen, C. Y., Lane, K., Chin, C., Lu, J., Kirsch, D. G., Golub, T. R., & Jacks, T. (2009). Dicer1 functions as a haploinsufficient tumor suppressor. *Genes Dev*, 23(23), 2700–2704.

[70] Laird, A. K. (1964). Dynamics of tumour growth. *British journal of cancer*, 18(3), 490.

[71] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. a., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*.

[72] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. a., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. a., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. a., Stegmaier, K., Bass, A. J., Garraway, L. a., Meyerson, M., Golub, T. R., Gordenin, D. a., Sunyaev, S., Lander, E. S., & Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), 214–8.

[73] Lee, A. J. X. & Swanton, C. (2012). Tumour heterogeneity and drug resistance: personalising cancer medicine through functional genomics. *Biochem Pharm*, 83(8), 1013–20.

[74] Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., DiPersio,

J. F., & Wilson, R. K. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218), 66–72.

[75] Loeb, L. a., Bielas, J. H., & Beckman, R. a. (2008). Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res*, 68(10), 3551–7; discussion 3557.

[76] Lopez-Garcia, C., Klein, A. M., Simons, B. D., & Winton, D. J. (2010). Intestinal stem cell replacement follows a pattern of neutral drift. *Science*, 330(6005), 822–5.

[77] Lynch, M. (2008). The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics*, 180(2), 933–43.

[78] Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA*, 107(3), 961–8.

[79] MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., Handsaker, R. E., Rosenfeld, J. A., Fromer, M., Jin, M., Mu, X. J., Khurana, E., Ye, K., Kay, M., Saunders, G. I., Suner, M.-M., Hunt, T., Barnes, I. H. A., Amid, C., Carvalho-Silva, D. R., Bignell, A. H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D. N., Xue, Y., Romero, I. G., Wang, J., Li, Y., Gibbs, R. A., McCarroll, S. A., Dermitzakis, E. T., Pritchard, J. K., Barrett, J. C., Harrow, J., Hurles, M. E., Gerstein, M. B., & Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070), 823–8.

[80] Mao, J.-H., Perez-losada, J., Wu, D., DelRosario, R., Tsunematsu, R., Nakayama, K. I., Brown, K., Bryson, S., & Balmain, A. (2004). Fbxw7/cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature*, 432(7018), 775–779.

[81] Martens, E. A., Kostadinov, R., Maley, C. C., & Hallatschek, O. (2011). Spatial structure increases the waiting time for cancer. *New journal of physics*, 13(11), 115014.

[82] May, R., Lawton, J., & Stork, N. (1995). *Assessing Extinction Rates*. Oxford University Press.

[83] McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., & Mirny, L. a. (2013). Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci USA*, 110(8), 2910–5.

[84] Merlo, L. M. F., Pepper, J. W., Reid, B. J., & Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6(12), 924–35.

[85] Merlo, L. M. F., Shah, N. a., Li, X., Blount, P. L., Vaughan, T. L., Reid, B. J., & Maley, C. C. (2010). A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev Res*, 3(11), 1388–97.

[86] Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4), R41.

[87] Meza, R., Jeon, J., Moolgavkar, S. H., & Luebeck, E. G. (2008). Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc Natl Acad Sci USA*, 105(42), 16284–9.

[88] Michor, F., Iwasa, Y., Lengauer, C., & Nowak, M. a. (2005). Dynamics of colorectal cancer. *Seminars in cancer biology*, 15(6), 484–93.

[89] Molloy, T., Devriese, L., Helgason, H., Bosma, A., Hauptmann, M., Voest, E., Schellens, J., & van't Veer, L. (2011). A multimarker qpcr-based platform for the detection of circulating tumour cells in patients with early-stage breast cancer. *British journal of cancer*, 104(12), 1913–1919.

[90] Moran, N. A. (1996). Accelerated evolution and muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences*, 93(7), 2873–2878.

[91] Moxon, E. R., Rainey, P. B., Nowak, M. a., & Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol*, 4(1), 24–33.

[92] Neher, R. a. & Shraiman, B. I. (2012). Fluctuations of fitness distributions and the rate of Muller's ratchet. *Genetics*, 191(4), 1283–93.

[93] Neznanov, N., Komarov, A. P., Neznanova, L., Stanhope-Baker, P., & Gudkov, A. V. (2011). Proteotoxic stress targeted therapy (pstt): induction of protein misfolding enhances the antitumor effect of the proteasome inhibitor bortezomib. *Oncotarget*, 2(3), 209.

[94] Nielsen, R. & Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, 20(8), 1231–9.

[95] Nowak, M. A. (2006). *Evolutionary dynamics: exploring the equations of life*. Belknap Press.

[96] Nowak, M. a., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I.-M., Vogelstein, B., & Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proc Natl Acad Sci USA*, 99(25), 16226–31.

[97] O'Brien, K. P., Remm, M., & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, 33(Database issue), D476–80.

[98] Page, D. L., Dupont, W. D., Rogers, L. W., & Rados, M. S. (1985). Atypical hyperplastic lesions of the female breast. A long-term follow-up study. *Cancer*, 55(11), 2698–708.

[99] Park, B.-J., Kang, J. W., Lee, S. W., Choi, S.-J., Shin, Y. K., Ahn, Y. H., Choi, Y. H., Choi, D., Lee, K. S., & Kim, S. (2005). The haploinsufficient tumor suppressor p18 upregulates p53 via interactions with atm/atr. *Cell*, 120(2), 209–221.

[100] Philipp-Staheli, J., Payne, S. R., & Kemp, C. J. (2001). p27< sup> kip1</sup>: Regulation and function of a haploinsufficient tumor suppressor and its misregulation in cancer. *Exp Cell Res*, 264(1), 148–168.

[101] Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. a., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. a., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A., & Stratton, M. R. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(14), 191–6.

[102] Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M.-l., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordoñez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A., & Campbell, P. J. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278), 184–90.

[103] Rossi-Fanelli, A., Cavaliere, R., Mondovì, B., & Moricca, G. (1977). *Selective Heat Sensitivity of Cancer Cells*. Berlin Heidelberg: Springer Berlin Heidelberg.

[104] S Datta, R., Gutteridge, A., Swanton, C., Maley, C. C., & Graham, T. a. (2013). Modelling the evolution of genetic instability during tumour progression. *Evol Appl*, 6(1), 20–33.

[105] Santagata, S. & Hu, R. (2011). High levels of nuclear heat-shock factor 1 (HSF1) are associated with poor prognosis in breast cancer. *Proceedings of the …*, 1.

[106] Segal, N. H., Parsons, D. W., Peggs, K. S., Velculescu, V., Kinzler, K. W., Vogelstein, B., & Allison, J. P. (2008). Epitope landscape in breast and colorectal cancer. *Cancer Res*, 68(3), 889–92.

[107] Sharma, S., Kelly, T. K., & Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, 31(1), 27–36.

[108] Sheltzer, J. M. & Amon, A. (2011). The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet*, 27(11), 446–53.

[109] Shibata, D., Peinado, M. A., Ionov, Y., Malkhosyan, S., & Perucho, M. (1994). Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nat Genet*, 6(3), 273–81.

[110] Silva, J. M., Garcia, J. M., Dominguez, G., Silva, J., Rodriguez, R., Portero, J. L., Corbacho, C., Provencio, M., España, P., & Bonilla, F. (2000). DNA damage after chemotherapy correlates with tumor response and survival in small cell lung cancer patients. *Mutat Res*, 456(1-2), 65–71.

[111] Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., & Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797), 268–74.

[112] Soong, N., Hinton, D., Cortopassi, G., & Arnheim, N. (1992). Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain. *Nature genetics*.

[113] Spencer, S. L., Gerety, R. A., Pienta, K. J., & Forrest, S. (2006). Modeling somatic evolution in tumorigenesis. *PLoS computational biology*, 2(8), e108.

[114] Stamatoyannopoulos, J. a., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., & Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nat Genet*, 41(4), 393–5.

[115] Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K. W., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M. S., Teague, J., Easton, D., Langerø d, A., Lee, M. T. M., Shen, C.-Y., Tee, B. T. K., Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.-F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., van 't Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. a. J. R., Salomon, A. V., Bø rresen Dale, A.-L., Richardson, A. L., Campbell, P. J., Futreal, P. A., & Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403), 400–4.

[116] Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng*, 12(5), 387–94.

[117] Tong, a. H., Evangelista, M., Parsons, a. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., Andrews, B., Tyers, M., & Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550), 2364–8.

[118] Tong, W. W. Y., Jin, F., McHugh, L. C., Maher, T., Sinclair, B., Grulich, A. E., Hillman, R. J., & Carr, A. (2013). Progression to and spontaneous regression of high-grade anal squamous intraepithelial lesions in HIV-infected and uninfected men. *AIDS*, 27(14), 2233–43.

[119] Vermeulen, L., Morrissey, E., van der Heijden, M., Nicholson, A. M., Sottoriva, A., Buczacki, S., Kemp, R., Tavaré, S., & Winton, D. J. (2013). Defining stem cell dynamics in models of intestinal tumor initiation. *Science*, 342(6161), 995–8.

[120] Vesely, M. D. & Schreiber, R. D. (2013). Cancer immunoediting: antigens, mechanisms, and implications to cancer immunotherapy. *Ann N Y Acad Sci*, 1284, 1–5.

[121] Wang, Z. C., Birkbak, N. J., Culhane, A. C., Drapkin, R., Fatima, A., Tian, R., Schwede, M., Alsop, K., Daniels, K. E., Piao, H., Liu, J., Etemadmoghadam,

D., Miron, A., Salvesen, H. B., Mitchell, G., DeFazio, A., Quackenbush, J., Berkowitz, R. S., Iglehart, J. D., Bowtell, D. D. L., & Matulonis, U. a. (2012). Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome. *Clin Cancer Res*, 18(20), 5806–15.

[122] Watanabe, S., Okita, K., Harada, T., Kodama, T., Numa, Y., Takemoto, T., & Takahashi, T. (1983). Morphologic studies of the liver cell dysplasia. *Cancer*, 51(12), 2197–2205.

[123] Wellings, S. R., Jensen, H. M., & Marcum, R. G. (1975). An atlas of subgross pathology of the human breast with special reference to possible precancerous lesions. *J Natl Cancer Inst*, 55(2), 231–73.

[124] Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., & Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, 308(5726), 1310–4.

[125] Wust, P., Hildebrandt, B., & Sreenivasa, G. (2002). Hyperthermia in combined treatment of cancer. *The lancet oncology*, (pp. 487–497).

[126] Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. a., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. a. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319), 1114–7.

[127] Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., & Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, 45(10), 1134–1140.

[128] Zhang, Q., Lambert, G., Liao, D., Kim, H., Robin, K., Tung, C.-k., Pourmand, N., & Austin, R. H. (2011). Acceleration of emergence of bacterial antibiotic resistance in connected microenvironments. *Science*, 333(6050), 1764–7.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.